

# EVALUATION AND ASSESSMENT

**Paul J. Black**

*King's College London, U. K .*

## 1 INTRODUCTION

### The purposes

Assessment in education has three main functions. The first is to record the achievements of individual pupils for the purpose of certification. The second is to record the achievements of groups, classes or schools, for broader policy purposes. The third is to serve teaching and learning.

The first function produces records which are passports - to better jobs or to higher education for a pupil leaving school. To fulfil this function, assessment has to command public confidence. In such assessment, there is also an aim of appraising a pupil's work as a whole, so that it can be described as summative.

The second function is characterised by emphasis on the public accountability both of individual schools, and of an education system at national or state level. The aim here is to inform policy by collection and analysis of evaluative information. Various regional and national monitoring systems, and international comparative studies, serve this purpose.

The third function arises because any learning system needs feedback. To serve this purpose, the assessment information has to provide information about each pupil's learning on the basis of which action can be taken to meet each pupil's learning needs. Such assessment may be called formative or diagnostic.

Ideally, each of these three functions requires assessment information of a different type from the other two. In practice, it is often necessary to use the same information to serve the different functions. Such multiple usage is attractive because it is economical, but it is not always feasible and there is always tension between the needs of the different functions.

### Social and political context

Differences in history, in traditions and in social and political needs have combined to produce systems which reveal on comparison, very different patterns of practice between different countries (Black 1992, Britton and Raizen 1996). For example, in some countries school-leaving certificates are entirely in the hands of each school, whilst in other countries such trust in teachers would be unthinkable and certificates are based entirely on external tests.

Such differences have developed historically. Social movements aimed at producing a more egalitarian society have in the past found ways to offset the advantages of privileged schools by use of assessment procedures. In such a venture, the external nature of the assessment system, and its lack of organic connection with the teaching system in schools is an essential feature. Conversely, in societies where equality of provision and status between schools has been achieved, it becomes possible to move from external and objective systems and to rely on each school's own role in assessment. However, a well-developed public and universal system of education becomes a very large user of public funds so that the justification of its budget claims becomes an important political issue. This pressure has led to demands for monitoring and accountability.

Such features affect the ways in which external examinations constrain the development of teaching and learning. Against complaints that they stifle and distort learning must be set the fact that, properly developed and applied, they can be powerful agents of reform.

The place of physics in the "high-stakes" certification assessments linked to university entrance is likewise very variable. In the university procedures of most countries, performance in several subjects, rather than in physics alone, is taken into account even for admission to specialised physics courses. Where there is a broad range of subjects taken at the top end of secondary school, there can be rules of aggregation which so operate that a student can be admitted to study in physics when the physics performance has been very low - as in Brazil. Where the entry examination is based on only a few subjects - as in the UK - it is possible to place great weight on the physics result, partly because this can be examined by an extensive exercise which would be impracticable for candidates taking more subjects.

### Structure of this chapter

This chapter will discuss in turn the three main functions, i.e.. for certification, for accountability and to serve learning. Inevitably, in consideration of the first of these, many general issues which apply to all three will have to be explored. A final section will review the interactions between these functions in attempting an overview of state and national systems.

## 2 ASSESSMENT FOR CERTIFICATION

### Methods - written tests

Paper and pencil tests have long been used as the main media in assessment in education. Conventional written tests in physics comprise a collection of items requiring definitions, standard explanations or derivations, accounts of standard experiments or applications, and a few problems, usually of a routine nature. The following example illustrates the combination :

- A sample of a Thallium element  $^{207}_{81}\text{Tl}$  of 16 gm emits beta radiation and is converted into an isotope of lead (Pb). The half-life of the decay process is 5 minutes, answer the following questions:

- What is meant by each of the following terms: isotopes - isobars - isotones ?

Give examples for each.

- Explain what is meant by "a half-life of 5 minutes for such a decay"

- Find the mass of the Thallium element  $^{207}_{81}\text{Tl}$  left in the sample after 1/3 hour.

- Calculate the decay constant of the Thallium element  $^{207}_{81}\text{Tl}$

- State (without explanation) two methods for radioactive waste disposal.

(Egypt - Wassef pp.57-68 in Black 1992)

Tests may be composed of shorter questions, treating the knowledge and the problem-solving aspects separately. The following is a typical example of a short problem style :

A compressed spring pushes apart two trolleys having masses of 0,2 kg and 0,3 kg, respectively, in such a way that the trolleys that were initially at rest are separated by 60 cm in 5 s. The mass of the spring and friction are negligible. What is the velocity of the trolleys? (Hungary - Radnai pp.84-100 in Black 1992)

Such short problem items need not necessarily be quantitative, as in this example : -

How does an electric dipole behave both in a uniform electric field and in a radial electric field? (Poland - Plazak and Mazur pp.125-139 in Black 1992)

More test items per unit time, and enhanced reliability in marking, may be secured with use of multiple choice items, which test problem solving abilities as well as knowledge, as in the following example : -

The points O, P, Q, R, S lie in a vertical line with equal intervals between them. An object is released from O with no initial velocity to fall freely.

The average velocity between O and T is equal to the instantaneous velocity at a point which is in the following interval : -

(a) between O and P; (b) between P and Q; (c) between Q and R; (d) between R and S; (e) between S and T.

(Japan - Ryu pp.101-123 in Black 1992)

Because of the large number of such items which pupils can attempt in a given testing time, such questions can achieve greater coverage and greater overall reliability. Their disadvantage is that they can give no direct evidence of a pupil's reasons for their choices, and some studies have shown that up to a third of pupils who choose the correct response may do so for a wrong reason (Tamir 1990). Exclusive use of them can lead to neglect, in teaching, of the discussion and argument which is a valuable and valid aspects of scientific enquiry. In public examinations, they have long been dominant in testing in the USA, are used as a minority component in some countries (e.g. UK, Sweden) and are not used at all in others (e.g. France)

All of the above examples involve a combination of knowledge of physics combined with skills of selecting and applying that knowledge. Where longer questions have a structure - as in the first example, that structure helps the respondent, both by presenting the selection of the knowledge needed for the problem, and by giving guidance for the sequence strategy needed to tackle it.

It is not necessary to rely entirely on remembered knowledge. Some written tests present a pupil with information in the form of a short article about a physics topic, and then assess the capacity to understand and apply by questions about it - either short open questions or multiple choice questions (see Black 1992)

### Assessing "process skills"

There have been attempts to assess "skills" separately from their physics knowledge. The problems that this raises can be illustrated by practical tasks set out to assess measurement skills. One problem is that of context - many pupils who seem able to use measuring instruments and procedures when required directly in artificial and isolated contexts do not deploy this ability when asked to undertake an investigation, even when provided with measuring instruments; they use only a qualitative comparison even when they have shown, in a different context, the ability to use the instruments. It is essential here to formulate a more comprehensive view of the skill involved. The ability to read accurately off a scale and to set up or adjust instruments is not enough. A scientist has to be clear about what to measure - for example measuring 'rate of flow' requires the understanding that a co-ordinated pair of measurements is needed. Furthermore, the investigator has to judge when to measure, which requires a judgement that the powerful tools of quantification can and should be applied to the problem (Black 1990).

Thus an assessment of skill with instruments and scales is not, on its own, of any great value, because such skill can only be useful if deployed in the light of a conceptual model of the system under investigation and of the variables involved. The same is true for other "skills" - for example observation, which is not passive reception but an essentially selective activity guided by assumptions about what should be selected. It follows that questions which test specific "skills" in isolation may not convey any useful information about a pupil's ability to use the skills in scientific work.

### Other methods - oral and practical

Oral tests are an important component of public examinations in physics in several Eastern European countries, but elsewhere they are seldom used. In the assessment of practical work, the logistic and other problems of using equipment in assessment exercises have led to attempts to use paper and pencil tests of practical work as substitutes. Such an approach can have undesirable feedback effects on teaching, and correlations between pupils' performances with real equipment and their responses to "equivalent" written tests are low. However, very few countries, notably the UK and Israel, use direct tests with equipment as part of public examination procedures (Britton and Raizen 1996).

One common form of practical test has been to specify set routines with given equipment and require certain types of response in terms of measurements taken and analysis of results. The tight constraints thus imposed improve reliability, but such skills as experimental design and choice of equipment are neglected. The "experiment" is indeed a vehicle for testing certain specific skills relevant to experimental science. An alternative approach has been to organise very short exercises, each testing a specific skill in its own set context. Examples of skills tested are the taking of accurate measurements with pre-set instruments, and making and recording qualitative observations of an unusual phenomenon.

Concern that out-of-context tests can be misleading has led to attempts to assess with open-ended and complete experimental tasks. If comparability and imposition of conventional examination conditions are required, then the best that can be done must be very limited. An additional constraint of such conditions is that, because all members of a group have to work simultaneously over a limited time, assessment has to be based on pupils' written records. Their actual actions, and their reasons for what they did, cannot be observed or interrogated directly. Also, since no system could handle more than a very few (probably one or two) such extensive tasks, generalisability of the results is a severe problem (Shavelson et al. 1993).

Such severe limitations can be overcome by assessment of extended pieces of work undertaken under normal working conditions. A written record of such work may hide some of the very important aspects of capability that a pupil may have exercised, which could only be appreciated if the process by which the pupil's work proceeded were observed and understood. The only feasible solution to such problems is for the teacher to be involved as assessor, allowing for such important features as the precise way the problem was posed, the effects of group collaboration and the constraints within which the pupil

had to work and the reasons for a pupil's decisions about his or her strategy. Such attention requires careful training of teachers, and time and opportunity for them to observe individual or group work carefully.

Of course, there is a paradox here. As the aims of physics education become more real and less artificial, they lead to activities which reflect more of the complexity and untidiness of real life, and so become less susceptible to any reproducible assessment process.

Such difficulties do not apply only to assessment of practical work. Some of the many types of written tasks that are in use are severely constrained by the time constraints of external test routines. If discussion of ideas, and the use of these, suitably selected and interwoven to bear on a new issue, is an important aim, then test procedures must work with responses in prose produced after adequate time for thought and planning.

### Combining methods

All methods of assessment suffer from shortcomings, and the best choice may often depend on the particular educational or social context involved. For example, availability of equipment, or the possibility of adequate monitoring to prevent unfair practices, may vary greatly from one situation to another. It is also necessary to consider that a combination of methods may be needed, both to reflect a range of aims, to enhance reliability and to offset the bias effects that any one technique is bound to introduce.

The range of methods used is very varied in some countries and very restricted in others. At one extreme is the exclusive use of multiple choice questions, as in the U.S.A. for those using one of the testing agencies, whilst at the other is the spectrum of eight types of task included in the Nuffield examination in the United Kingdom (Black 1992). If there is a most popular pattern, it seems to be a mixture of multiple choice questions and short problems.

One reason why such variety is found is that familiarity with certain methods and a tradition of using them can inhibit the possibility at looking seriously at alternatives. For example, there is a tradition in some countries that the mathematical theoretical question is the most significant test of ability as a physicist, so these are given far greater prominence than, say, ability to tackle problems experimentally or to write critically about the subject.

Another reason may be a belief that some methods cannot give reliable results : some think that multiple choice questions, with their objective marking, pre-testing checks, and possibilities of statistical analyses of large numbers of questions, give results which are so much more reliable than other types that they should be the only method used.

A third reason may be cost. For examinations with a large entry, the expense of preparing good multiple choice questions is justified by the possibility of securing reliable marking at low cost. To set and mark several other types of question makes heavy demands on the time and expertise of examiners. In particular, assessment of practical work is usually ruled out on grounds of cost and practicability.

### Reliability

The question of whether or not the range of types of questions really matters is a more complex one to address. The main factors involved are reliability, validity, feedback, quality and bias. Reliability is the simplest to tackle. An examination can be reliable if one can have confidence that the same results would be obtained with a parallel examination, i.e. one set and marked according to the same aims and methods. It is possible to obtain a measure of reliability by testing the internal consistency of responses, but this is only appropriate where there is a reasonably large number of questions and where it can be assumed that they should give homogeneous results. A more strict test is to set parallel forms to the same students, or at least to give some students large numbers of questions in the same domain of testing and to determine from the results the minimum number of questions needed to reduce the error below a given limit. Such tests are hardly ever carried out - the faith of university and school examiners in the reliability of short external test papers is usually unjustified in that it is not based on evidence. Reliability of marking is also an issue in public examinations - careful training of examiners and cross-checks on their marking are essential.

### Validity

A broad concept of validity is both necessary and cogent in any attempt to improve the quality of assessments. The opening statement of Messick's (1989) review gave an authoritative definition :

*Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment.*

Thus, if a test is meant to show whether or not a pupil is competent in measurement, it is a matter for expert judgement whether it calls for use of those abilities in measurement which are of value in science. If a test is designed to help predict potential for future attainment, this can be appraised, after the event, by studying correlation of its results with those which it was designed to predict. 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. In this light, classroom assessment has a better chance of success than formal timed written tests.

### Effects on learning

One focus of concern about testing is its effect on learning. Research is showing that preparing pupils for multiple choice tests can be inimical to good learning practice. This is expressed by Resnick and Resnick (1992) as follows : -

*Children who practice reading mainly in the form in which it appears in the tests - and there is good evidence that this is what happens in many classrooms - would have little exposure to the demands and reasoning possibilities of the thinking curriculum. . . .*

*Students who practised mathematics in the form found in the standardised tests would never be exposed to the kind of mathematical thinking sought by all who are concerned with reforming mathematical education . . .*

*Assessments must be so designed that when you do the natural thing - that is, prepare the students to perform well - they will exercise the kinds of abilities and develop the kinds of skills that are the real goals of educational reform.*

### Quality

Of the various features that affect the quality of testing and assessment, one that stands out here must be about the time taken for external tests. If one country invests over eight hours in tests in one subject to determine future career prospects, how can another perform the same function in under two hours ? Any test is meant to sample the several domains of performance that are important in the subject - the A.P.U. Science monitoring at Age 13 in 1984 used 35 one-hour tests to obtain reliable results over the domains judged important for science (Johnson 1988).

One tactic to overcome the difficulty is to narrow the range of the domains assessed, but this narrows the scope of subject aims which the examinations exhibit and thereby, because of their high stakes in most systems, the aims that they impose on school learning.

A related issue pertaining to quality lies in the balance between questions which can be answered by routine procedures using learned algorithms and those which require thoughtful translation and application of principles and procedures. In many country's systems, the analysis shows that the balance is very much in favour of the former, and this must in part be due to the limitations of test times. The very difficult task for most test setters is to develop questions which demand thoughtfulness rather than rote learning, at a level where the average pupils can have a good chance of success.

### **Bias**

Any assessment process is an interaction between certain questions, items and/or procedures, and the pupil being assessed. There are many ways in which this interaction may operate defectively, so producing bias or flaws in the results (Gipps and Murphy 1994).

One example that has been well researched is gender bias. An issue of comparable importance is the problem of cultural or ethnic bias. Difficulties arise in the language used in teaching the subjects, in the range and nature of "everyday" examples used and in cultural assumptions inherent in western science.

There are also many ways in which individual pupils even within the same gender and cultural group, can be unfairly affected by the presentation, context, or language, of assessments. Some pupils who produce strange and apparently worthless responses can, in discussions reveal that these were sensible responses to a misinterpretation of the question (Gauld 1980).

### **Norm and criterion referencing, aggregation and profiles**

The dominant tradition in education at all levels has been to use norm-referenced tests. Where these are used, the emphasis is on comparison of any one pupil or group with the whole population entered for the assessment or test. The alternative approach is to give priority to criterion referencing, so that an assessment result implies that a pupil has satisfied certain defined criteria irrespective of the achievements of other pupils (Gipps 1994).

A related issue is that of the agglomeration of results. A common approach is to take results of marking questions, on different topics and assessing different abilities, and to sum these to give each pupil a total score. This approach can be compared to attempts to obtain a composite measure of chalk, cheese and the smell of a rose. One solution is to report results as a profile in which separate elements are kept apart. It may then be possible to say that scores on each component have meaning in relation to success on given criteria.

One outcome of such discussion is to cast doubt on whether the conventional practice, of setting tests with questions ranging over a variety of contexts, knowledge and skills and adding marks to produce a single norm-referenced number, is valid i.e. can give meaningful results and can be justified in relation to uses made of them.

### **Teachers' assessments**

In principle, a teacher who can record a pupil's performance over time and in several contexts, and who can discuss idiosyncratic answers in order to understand the thinking that might lie behind them, can build up a record of far better reliability than any external test can achieve. Similarly, it also seems that the need to secure validity by assessing pupils skills in the context of tackling realistic problems, and to do this over several different problem contexts, can only be met by teachers' assessments.

The obvious difficulties are the lack of teachers' assessment expertise, the lack of comparability of standards between different schools, and indeed between teachers in the same school, and the dangers of prejudice and dishonesty. None of these obstacles can be overcome quickly or cheaply.

Comparability between schools can be secured by exchange meetings between groups of schools in which criteria are discussed and examples of student work are exchanged. 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. There are other methods of moderating standards between schools, notably using external visitors to inspect procedures and submitted samples of work (Black 1993).

The UK, seems unique giving weight in public certificate examinations to marking by the students' own teachers of work done outside formal examination conditions. For this to be acceptable, in commanding public confidence, there have to be very carefully specified rules, and a system which, by drawing samples from schools and checking their standards, can ensure the integrity, and comparability of the results. Sweden is also unique in a very different way. There, the teachers' assessments are the foremost element in determining the whole result. The external examination serves to calibrate the schools distribution as a whole, but leaves the teacher free to decide about individual students (Black 1992).

## **3 ASSESSMENT FOR ACCOUNTABILITY**

The aim involved here is to inform public policy by collection and use of evaluative information. Where public certification results are accumulated for use as performance indicators of schools, the data lacks necessary detail. However, for this purpose there is no need to produce reliable and complete results for any one individual, different pupils can be given different tests and performance can therefore be investigated in much greater detail.

If relationships between performance outcomes, and other factors that can be adjusted by public or school policy, are to be explored, information on these factors has to be collected. Thus the task involves selection and collection of data on such features as class size, home background of pupils, time spent on learning, laboratory equipment, and so on, to supplement the pupil performance data. The analysis of possible relationships within the data then becomes complex because multiple correlations have to be carefully explored. Any one factor (e.g. type of school) can be correlated with, and so appear to cause, performance variations when it is really only a surrogate for a different factor (e.g. attainment of pupil on entry to the school) with which it is associated. The interpretation of relationships is also difficult because correlations even when properly isolated by statistical means cannot by themselves offer proof of causes (Woodhouse and Goldstein 1996).

Any national monitoring has to reflect a structure of aims and criteria. Insofar as it produces good quality test items and comprehensive data about these aims and criteria, it will attract the attention of teachers and influence their work. Thus, in the national science monitoring in the UK, the structure chosen emphasised process aims in science and influenced teachers to give more emphasis to such aims as observation and the design of practical investigations (Black 1990). Thus the monitoring became prescriptive, and it could be said that the exercise was as much a curriculum development project as an assessment project because it moved ahead of current practice. Thus monitoring can provide an opportunity to promote innovation, just as it can be a powerful conservative force if it only reflects established aims and procedures.

## **4 SUPPORT FOR LEARNING - FORMATIVE ASSESSMENT**

## Introduction to the issues

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.

Several common features emerge from surveys research into formative assessment in many countries (Black, P.J. 1993). One of these is that there is substantial evidence that carefully designed programmes of formative assessment do lead to improvements in pupils' learning. A common characteristic is that the renewal of assessment practice is part of a wider change in teaching strategy and not merely an added supplement.

Another feature is that assessment generally, and formative assessment in particular, ranks low in teachers own practice and priorities. Furthermore, 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. Recent surveys have also reported that there was very little formative assessment to be seen in science work (Russell et al. 1994).

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. This has many damaging consequences. External tests create models and images for the whole of assessment and testing which are misleading. For example, a common teaching practice is to set an end-of-unit or end-of-term test which resembles the external tests as closely as possible, and to record the results, perhaps even make them public. Since the data are not used to modify teaching and learning, there is no formative assessment. Thus a practice of more or less frequent summative assessment is set up, so that assessment is equated with testing and acquires the negative overtones, of ordeal by fire for pupils, and of onerous and unproductive labour for their teachers.

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.

## Two examples

Two specific examples, each about the development of practice in a school in England will serve to illustrate some of the issues. In the first school (Parkin and Richards 1995), the science teachers wanted to use pupils' self assessment and subsequent teacher/pupil discussion as the basis for their assessments. For each module of the course, target criteria were expressed in language accessible to pupils. For each lesson, every pupil had a sheet setting out the criteria with a space opposite each in which the pupil had to state whether the criterion was clear and had been achieved; pupils were also asked to write in other comments - for example about enjoyment or interest.

The teacher subsequently annotated each of the criterion responses with one of three code letters as follows : **A** - for full understanding achieved, **P** - for a partial understanding achieved, **V** - where the work had been no more than 'visited' by the pupil.

It took about a year from the first introduction of this scheme for pupils to learn to use it productively- many at the start wrote very few, and very vague, comments, but during the year these change and become more explicit and perceptive and so more useful. Pupils were not accustomed to thinking of their own learning as purposeful in relation to target criteria. They were also had to break away from treating the assessment as a formal testing.

Some pupils, especially the less able, did not like to admit failure and sometimes said that they understood when they did not. Teachers tried emphasise to each pupil that the record was a private document aimed to help the teacher to see the pupil's problems so that help could be given where needed.

In the second school (Fairbrother 1995), a teacher of physics to a class of 12/13 year-olds wanted them to approach a unit on electricity and magnetism in a more responsible way. He aimed to help them to

put each lesson into the context of the whole unit,

have a summary of what they had been doing for revision purposes,

see what was to come next in the unit of work.

He gave each pupil a "revision sheet" for the unit containing about 25 target statements, for example : -

*Know how to make an electromagnet and how to vary its strength*

*Know that a complete circuit is needed for electrical devices to work*

*Know that a wire carrying a current in a magnetic field will try to move*

*Know how switches, relays, variable resistors, sensors and logic gates can be used to solve simple problems e.g.. burglar alarm, frost warning, automatic street lights*

Most of the pupils had little idea of how to use this list, for example to check the notes in their exercise book against its contents, or to check whether they did know what was required. Some of the less-organised pupils simply lost it, others simply stored it away and were not referring to it.

The teacher's explanation of this failure was that these pupils were being given too much teaching only about the subject and not about how to learn. The revision sheet was intended to address this issue but the teacher had not realised at the beginning how much actual teaching on how to use the sheet would be needed. For example, when pupils were told as a homework to "revise for the test", most of them were floundering. There seemed to be two main reasons for this. The first was that pupils do not know how to extract from everything they do that which they are supposed to know and understand. Teachers know the difference between the *ends* which they want to achieve and the *means* by which they are trying to achieve them. The pupils do not see this difference. A second reason was that pupils did not know how they would be expected to show their knowledge and understanding. Most of them learn by experience something of what is required of them, and for many pupils this experience is hard and dispiriting. Some of them, usually the weakest,

never do learn.

### Developing Good Practice

The traditional dominance of the summative function means that formative assessment struggles for its status and development (Fairbrother et al. 1995). 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 function can inhibit the growth of the formative function in teachers' assessments in several ways. Summative practice can mislead 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 (e.g. 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 judgements 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", formative practice, with its priorities of identifying and helping to meet the learning needs of each pupil, can treat different pupils very differently;
- summative purposes can demand documented evidence for results - e.g. 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;

An outstanding source of difficulty in developing formative assessment is that it cannot just be stuck on to existing schemes of work, it has to be built into 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 organisation that such use requires.

Effective use of assessment feedback requires teacher judgement, 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.

Use of formative evidence is perhaps the most challenging aspect. There are 'macro' responses, in terms of streaming and setting, but these do not deal with immediate needs. Some teachers have responded by organising 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, H. 1993).. Others indicate less formal and more flexible approaches, building in revision or re-visiting opportunities in later work for those in need. Affecting this last issue is the extent to which teaching programmes are flexible rather than rigid.

The technology of collecting data on pupils' progress is only just beginning to develop. 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. The sheets described in the first example above show one way to do this; the outcomes are distinctive in that they produce detailed information in relation to statements about specific aims - i.e. they are quite naturally implementing criterion referencing because this is what formative assessment inevitably needs. Furthermore, because these provide written evidence in a systematic way, they relieve the teacher from the pressure of noting and recording entirely from the ephemeral evidence of classroom events. Such ephemeral evidence can however have its own unique value: some have found it especially useful - and surprising - 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 pupils (see Cavendish et al. 1990, Connor 1991).

When an assessment activity is so closely built into a learning programme, it would be foolish to prevent pupils from commenting on their results, from challenging them, and from repeating assessments if they so wished to improve their performance. Thus formative assessments become both informal, and pupil driven, as a consequence of their role in supporting learning. The prominence given to pupils' self-assessment is a notable feature and experience shows that pupils cannot play an effective part in their own assessment except within a long-term programme designed to help them achieve and sustain an overview of their learning targets and to apply the criteria which comprise it to their own progress. As both of the examples above make clear, pupils have to be taught how to assess their own progress. An important part of this work is the translation of curriculum aims into language that all pupils can understand, and down to a level of detail that helps them relate directly to their learning efforts. It also follows that targets have to be both attainable in the short term, and adequately modest in relation to the learners' prospects of success. These requirements come out in particularly sharp forms in providing for pupils with special learning difficulties - but they are important for all.

Teachers who have developed pupils' self-assessment report on several advantages which follow - that pupils can direct their own efforts more clearly and effectively, that they can be more actively involved and motivated in relation to their own learning progress, that they can then suggest their own ways to improve their attainment, and even that they can challenge assessments which they believe to be unfair.

Clearly, pupils' involvement can make it more feasible for teachers to carry through a programme of formative assessment. However, this involvement also changes both the role of the pupil as learner and the nature of the relationship between teacher and pupil, making the latter shoulder more of the responsibility for learning. Quite apart from the needs for improved assessment, the prior need for improved learning demands such changes. Indeed, some have argued that meta-cognition, by which they mean awareness and self-direction about the nature of their learning work, is essential to pupils' development in concept learning, and the work described here is clearly serving that purpose (see Brown 1987, White and Gunstone 1989, Baird and

Northfield 1992). Thus improved formative assessment can lead to changes which are of much wider significance - changes which should be a powerful help with pupils' personal development and which should also be part of any programme to help them to be more effective learners.

## 5 SYSTEMS AND ROLES

Good summative assessment requires the involvement of teachers, so there seems to be no alternative to developing ways in which teachers can carry both summative and formative roles, using at least some, but not necessarily all, of their evidence for both but distinguishing carefully between the methods and needs which relate to the two purposes. Carrying two roles in this way would be very demanding. On the one side, there are the learning needs of their pupils, which it must be their first concern to serve. On the other side are the 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. The teacher has to hold the boundary between the different pressures coming from these two sides.

The main reason for emphasising this issue is that some of the most important aims of physics education cannot be reflected in, and so supported by, assessment systems which rely only on short external testing. Reform of national summative testing is a serious necessity. In 1992, at the end of a review of national tests in physics in eleven countries, I wrote the following in a closing summary :-

*One conclusion that I draw from this study is that the range of methods used, and the range of abilities assessed, by these physics examinations is too narrow and that they are probably having a seriously narrowing effect on the development of school physics and on the recruitment of physicists. There are many reasons for this. Shortage of resources, together with other system constraints, may account for the willingness of physics examiners to work with systems which they judge to be, at best, far from ideal and perhaps seriously damaging to the future of physics.. Perhaps this situation is accepted too readily by all of us.*

(Black 1992)

Public examinations have particular power over the future of physics. By setting the targets and framework within which high school teachers feel they must work, they determine the structure and the image of the subject in the eyes of the young. If such examinations do not call for or promote activities which are important and attractive to physicists and if they convey a very narrow image of the subject, they will attract too few specialists, and give all adults a negative view of physics.

## References

- Baird, J.R. and Northfield, J.R. (eds.) (1992) *Learning from the PEEL experience*. Melbourne: Monash University.
- Black, H (1993) Assessment: A Scottish Model pps.91-94 in Fairbrother, R., Black, P.J. and Gill, P. (eds.) *TAPAS : Teacher Assessment of Pupils: Active Support*. King's Education Papers No.3. London: C.E.S. King's College.
- Black, P.J. (1990) APU Science - the past and the future. *School Science Review* 72. 13-28
- Black, P.J. (1992) *Physics Examinations for University Entrance : an International Study*. Science and Technology Education - Document No. 45. Paris : UNESCO.
- Black, P.J. (1993), Formative and Summative Assessment by Teachers. *Studies in Science Education*. 21. 49 - 97.
- Britton, E.D. and Raizen, S.A. (eds.) (1996) *Examining the Examinations : An International Comparison of Science and Mathematics Examinations for College-Bound Students*. Boston : Kluwer
- Brown, A. (1987) Metacognition, executive control, self-regulation and other mysterious mechanisms. pps 65 - 116 in Weinert, F.E and Kluwe, R.H. (eds.) *Metacognition, Motivation, and Understanding*. Hillsdale, New Jersey: Lawrence Erlbaum.
- Cavendish, S., Galton, M., Hargreaves, L. and Harlen, W. (1990) *Observing Activities*. London: Paul Chapman.
- Connor, C. (1991) *Assessment and Testing in the Primary School*. London: Falmer Press.
- Fairbrother, R. (1995) *Pupils as Learners*. pp.105-124 in Fairbrother et al. op.cit.
- 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, .
- Gipps, C.V. & Murphy, P. (1994) *A fair test ? Assessment, achievement and equity*, Milton Keynes : Open University Press.
- Gauld, C.F. (1980) Subject oriented test construction, *Research in Science Education*, 10, 77--82.
- Johnson S. (1988) *National Assessment : the APU Science Approach*. London : Her Majesty's Stationery Office.
- Messick, S. (1989) Validity pp. 12 - 103 in Linn, R.L. (ed.), *Educational Measurement (3rd. Edition)*, London : Collier Macmillan.
- Parkin, C and Richards, N (1995) *Introducing Formative Assessment at KS3 : an attempt using pupils' self-assessment*. pp 13-28 in Fairbrother et al. op.cit.
- Resnick, L.B. & Resnick, D.P.(1992) Assessing the Thinking Curriculum: New Tools for Educational Reform pp. 37 - 75 in Gifford, B.R. & O'Connor, M.C.(eds.), *Changing Assessments : Alternative Views of Aptitude, Achievement and Instruction*, Boston : Kluwer.
- Russell, T., Qualter, A., McGuigan, L. and Hughes, A. (1994), *Evaluation of the implementation of Science in the National Curriculum at Key Stages 1, 2 and 3*. London: School Curriculum and Assessment Authority.
- Shavelson, R.J., Baxter, G.P. & Gao, X.(1993) Sampling variability of performance measurements *Journal of Educational Measurement* 30. 215--232.

Tamir, P. (1990) Justifying the selection of answers in multiple-choice questions. *International Journal of Science Education* 12. 563-573.

White, R.T. and Gunstone, R.F. (1989) Meta-learning and conceptual change. *International Journal of Science Education*. 11. 577-586.

Woodhouse, G. and Goldstein, H. (1996) The Statistical Analysis of Institution-based Data pp.135-144 in Goldstein, H. and Lewis, T. (eds.) *Assessment: Problems, Developments and Statistical Issues* Chichester UK : Wiley

\*\*\*\*\*

Section E2, *Evaluation and Assessment* from: *Connecting Research in Physics Education with Teacher Education*  
An I.C.P.E. Book © International Commission on Physics Education 1997,1998  
All rights reserved under International and Pan-American Copyright Conventions

[Return to the Table of Contents](#)