Inside the Black Box

Raising Standards Through Classroom Assessment

Paul Black and Dylan Wiliam

King's College London School of Education

The Black Box

Raising the standards of learning that are achieved through school education is an important national priority. Governments have been vigorous in the last ten years in making changes in pursuit of this aim. National curriculum testing, the development of the GCSE, league tables of school performance, initiatives to improve school planning and management, target setting, more frequent and thorough inspection; these are all means to the end. But the sum of all of these doesn't add up to an effective policy because something is missing.

Learning is driven by what teachers and pupils do in classrooms. Here, teachers have to manage complicated and demanding situations, channelling the personal, emotional and social pressures amongst a group of 30 or so youngsters in order to help them to learn now, and to become better learners in the future. Standards can only be raised if teachers can tackle this task more effectively—what is missing from the policies is any direct help with this task.

In terms of systems engineering, present policy seems to treat the classroom as a **black box**. Certain *inputs* from the outside are fed in or make demands—pupils, teachers, other resources, management rules and requirements, parental anxieties, tests with pressures to score highly, and so on. Some *outputs* follow, hopefully pupils who are more knowledgeable and competent, better test results, teachers who are more or less satisfied, and more or less exhausted. But what is happening inside? How can anyone be sure that a particular set of new inputs will produce better outputs if we don't at least study what happens inside?

The answer usually given is that it is up to teachers—they have to make the inside work better. This answer is not good enough for two reasons. First, it is at least possible that some changes in the inputs may be counter-productive—making it harder for teachers to raise standards. Secondly, it seems strange, even unfair, to leave the most difficult piece of the standards-raising task entirely to teachers. If there are possible ways in which policy makers and others can give direct help and support to the everyday classroom task of achieving better learning, then surely these ways ought to be pursued vigorously.

None of the reform items mentioned in the first paragraph is aimed at direct help and support. To be sure, inspections do look inside classrooms, and insofar as they focus on what is happening there they draw attention to important issues. But they are not designed to give help and support, recommendations being in very general terms.

This paper is about the inside of the black box. It is focused on one aspect of teaching—formative assessment, but the argument that we develop is that this feature is at the heart of effective teaching.

The argument

We start from the self-evident proposition that teaching and learning have to be interactive. Teachers need to know about their pupils' progress and difficulties with learning so that

they can adapt their work to meet their needs—needs which are often unpredictable and which vary from one pupil to another. Teachers can find out what they need in a variety of ways — from observation and discussion in the classroom, and from written work of pupils whether done as homework or in class.

In this paper, the term 'assessment' refers to all those activities undertaken by teachers, *and by their students in assessing themselves*, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged. *Such assessment becomes 'formative assessment' when the evidence is actually used to adapt the teaching work to meet the needs.*

There is nothing new about this. All teachers make assessments in every class they teach. But there are three important questions about this process which this paper sets out to answer. These are:

First **Is there evidence that improving formative assessment raises standards?**

Second **Is there evidence that there is room for improvement?**

Third **Is there evidence about how to improve formative assessment?**

In setting out to answer these questions, we have conducted an extensive survey of the research literature. This has involved checking through many books, through the issues of over 160 journals for the past nine years, and studying earlier reviews of research. This process yielded about 580 articles or chapters to study. Out of this we have prepared a lengthy review, which uses material from 250 of these sources. The review has been published in the journal "Assessment in Education" (Black and Wiliam, 1998) together with comments on our work by leading educational experts from Australia, France, Hong Kong, Southern Africa and the USA.

The conclusion we reach from the full review is that the answer to each of the above three questions is a clear 'Yes'. The three main sections of this present paper will outline the nature and force of the evidence which justifies this conclusion. However, we are presenting here a summary, so that our text will appear strong on assertions and weak on the details of their justification. Our position is that these assertions are all backed by evidence, and that this backing is set out in full detail in the lengthy review on which this present paper is based.

We believe that our three sections establish a strong case—a case that government, its agencies, and the teaching profession should study very carefully if they are seriously interested in raising standards in education. However, we also acknowledge widespread evidence that fundamental educational change can only be achieved slowly — through programmes of professional development that build on existing good practice. Thus, we are not concluding that, in formative assessment, we have yet another 'magic bullet' for education. The issues involved are too complex and too closely linked to both the difficulties of classroom practice and the beliefs that drive public policy. In a fourth and final section we confront this complexity and try to sketch out a strategy for acting on our evidence.

Is there evidence that improving formative assessment raises standards ?

A review published in 1986, concentrating—but not exclusively—on classroom assessment work for children with mild handicaps, surveyed a large number of innovations from which 23 were selected (Fuchs and Fuchs, 1986). This group all satisfied the condition that quantitative evidence of learning gains was obtained, both for those involved in the innovation, and for a similar group not so involved. Since then, many more papers have been published describing

similarly careful quantitative experiments. Our own review has selected at least 20 more such studies—the number depends on how rigorous a set of selection criteria are applied. All of these studies show that innovations which include strengthening the practice of formative assessment produce significant, and often substantial, learning gains. These studies range over ages (from 5-year olds to university undergraduates), across several school subjects, and over several countries.

For research purposes, learning gains of this type are measured by comparing (a) the average improvements in pupils' scores on tests with (b) the range of scores that are found for typical groups of pupils on these same tests. The ratio of (a) divided by (b) is known as the *effect size*. The formative assessment experiments produce typical *effect sizes* of between 0.4 and 0.7 : such effect sizes are larger than most of those found for educational interventions. The following examples illustrate some practical consequences of such large gains:

- An effect size of 0.4 would mean that the average pupil involved in an innovation would record the same achievement as a pupil just in the top 35% of those not so involved.
- A gain of effect size 0.4 would improve performances of pupils in GCSE by between one and two grades.
- A gain of effect size 0.7, if realised in the recent international comparative studies in mathematics (TIMSS—Beaton et al., 1996), would raise England from the middle of the 41 countries involved to being one of the top 5.

Some of these studies exhibit another important feature. *Many of them show that improved formative assessment helps the (so-called) low attainers more than the rest, and so reduces the spread of attainment whilst also raising it overall.* One very recent study is entirely devoted to low attaining students and students with learning disabilities, and shows that frequent assessment feedback helps both groups enhance their learning (Fuchs et al. 1997). Any gains for such pupils could be particularly important, for any 'tail' of low educational achievement is clearly a portent of wasted talent. Furthermore, pupils who come to see themselves as unable to learn usually cease to take school seriously—many of them will be disruptive within school, others will resort to truancy. Given the habits so developed, and the likelihood that they will leave school without adequate qualifications, such pupils are likely to be alienated from society and to become the sources and the victims of serious social problems.

So it seems clear that very significant learning gains could lie within our grasp. The fact that such gains have been achieved by a variety of methods which have, as a common feature, enhanced formative assessment indicates that it is this feature which accounts, at least in part, for the successes. However, it does not follow that it would be an easy matter to achieve such gains on a wide scale in normal classrooms. The reports which we have studied bring out, between and across them, other features which seem to characterise many of the studies, namely:

- All such work involves new ways to enhance feedback between those taught and the teacher, ways which require new modes of pedagogy—which will require significant changes in classroom practice.
- Underlying the various approaches are assumptions about what makes for effective learning—in particular that students have to be actively involved.
- For assessment to function formatively, the results have to be used to adjust teaching and learning—so a significant aspect of any programme will be the ways in which teachers do this.
- The ways in which assessment can affect the motivation and self-esteem of pupils, and the benefits of engaging pupils in self-assessment, both deserve careful attention.

Is there evidence that there is room for improvement ?

A poverty of practice

There is a wealth of research evidence that the everyday practice of assessment in classrooms is beset with problems and short-comings, as the following quotations indicate:

> "Marking is usually conscientious but often fails to offer guidance on how work can be improved. In a significant minority of cases, marking reinforces under-achievement and under-expectation by being too generous or unfocused. Information about pupil performance received by the teacher is insufficiently used to inform subsequent work."
>
> (*OFSTED general report on secondary schools 1996, p.40.*)
>
> "Why is the extent and nature of formative assessment in science so impoverished?"
>
> (*UK secondary science teachers—Daws and Singh, 1996 UK*)
>
> "The criteria used were 'virtually invalid by external standards'"
>
> (*French primary teachers—Grisay, 1991*)
>
> " Indeed they pay lip service to it but consider that its practice is unrealistic in the present educational context"
>
> (*Canadian secondary teachers—Dassa, Vazquez-Abad and Ajar, 1993*).

The most important difficulties, which are found in the UK, but also elsewhere, may be briefly summarised in three groups. The first is concerned with *effective learning* : -

- Teachers' tests encourage rote and superficial learning; this is seen even where teachers say they want to develop understanding—and many seem unaware of the inconsistency.
- The questions and other methods used are not discussed with or shared between teachers in the same school, and they are not critically reviewed in relation to what they actually assess.
- For primary teachers particularly, there is a tendency to emphasise quantity and presentation of work and to neglect its quality in relation to learning.

The second group is concerned with *negative impact* : -

- The giving of marks and the grading functions are over-emphasised, while the giving of useful advice and the learning function are under-emphasised.
- Use of approaches in which pupils are compared with one another, the prime purpose of which appears to them to be competition rather than personal improvement. In consequence, assessment feedback teaches pupils with low attainments that they lack 'ability', so they are de-motivated, believing that they are not able to learn.

The third group focuses on *the managerial role* of assessments

- Teachers' feedback to pupils often seems to serve social and managerial functions, often at the expense of the learning functions.
- Teachers are often able to predict pupils' results on external tests—because their own tests imitate them—but at the same time they know too little about their pupils' learning needs.
- The collection of marks to fill up records is given greater priority than the analysis of pupils' work to discern learning needs; furthermore, some teachers pay no attention to the assessment records of previous teachers of their pupils.

Of course, not all of these descriptions apply to all classrooms, and indeed there will be many schools and classrooms to which they do not apply at all. Nevertheless, these general conclusions have all been drawn by authors in several countries, including the UK, who have collected evidence by observation, interviews and questionnaires from many schools.

The empty commitment

The changes in England and Wales since the 1988 Education Reform Act have had powerful effects on assessment. The statements of policy which preceded that Act, the recommendations of the TGAT (Task Group on Assessment and Testing, DES 1988) report, and all subsequent statements of government policy, have emphasised the importance of formative assessment by teachers. However, most of the available resources, and public and political attention, have been concentrated on the tests which are given at the end of the Key Stages to yield overall levels or grades, and whilst teachers' contributions to these 'summative' assessments have been given some formal status, hardly any attention is paid to them. Moreover, the problems of the relationship between teachers' formative and their summative roles have received no attention.

There is indeed a very sharp contrast between these formal commitments, to the central importance of formative assessment, and the actual priority given to it. The clearest evidence of this is in the detailed account—written by one of its members—of the work of the Schools Examinations and Assessment Council between its foundation in 1988 and 1993 (Daugherty, 1995). During that time, teachers' assessments appeared as an explicit item on that Council's agenda on only two occasions, each time because the government department (then the Department of Education and Science) had addressed a specific question about summative aspects, whilst the formative aspects of teachers' assessments were never given serious attention. Thus the body charged to carry out government policy on assessment had no strategy either to study or to develop the formative assessment of teachers, and did no more than devote a tiny fraction of its resources to publications concerned with such work.

The political commitment to external testing of teachers and schools in order to promote competition through league tables had a central priority, whilst the commitment to formative assessment was probably a marginal feature. As researchers the world over have found, external tests, such as our the National Curriculum tests and the GCSEs, which function, to use the American phrase, as 'high stakes' tests, always dominate both teaching and assessment. In particular, because of their constraints and their function to provide overall summaries of achievement rather than helpful diagnosis, they give teachers poor models for formative assessment.

It is also possible that many of the commitments were stated in the belief that formative assessment was not problematic—that it already happened all the time and it needed no more than formal acknowledgement of its existence. Some attempts were made, by the School Examinations and Assessment Council (SEAC) and subsequently by its successor the School Curriculum and Assessment Authority (SCAA) to support teachers assessments by producing general guides to procedures, and by publishing examples of pupils' work with guidance on how these concrete examples would be assessed. The general guides were not found to be helpful, and they could not be, given that they were not based on serious study of practical problems. The materials for exemplification have been valuable, but being guides to the interpretation of national curriculum criteria in the marking of pupils' work, they do not constitute a significant contribution to the development of formative work, and indeed might enhance the summative rather than the formative roles of teachers' assessment work.

Given this, it is hardly surprising that numerous research studies of the implementation of the UK's educational reforms have found that formative assessment is, as one put it, "seriously in need of development" (Russell *et al.* 1995). However, more recent research studies have found some improvement in formative practice in primary schools (Gipps et al. 1996), and over the past two years, the DfEE have allocated in-service (GEST) funds to the specific purpose of developing teacher assessment at Key Stage Two, and this has made it possible for some LEAs

to begin to improve formative assessment through in-service training. Such developments are welcome, but as yet they do not begin to redress the effects of neglect and of lost opportunities.

With hindsight, it can be seen that the failure, to perceive the need for substantial support for formative assessment and to take responsibility for developing such support, was a serious error. Even in relation to the needs of the education system before 1988, formative assessment was weak. Given the new and mountainous burdens of the National Curriculum changes, it should have been clear that existing good practice could hardly have survived, let alone have risen to the challenge of a far more demanding set of requirements.

Is there evidence about how to improve formative assessment?

The self-esteem of pupils

> "... a number of pupils ... are content to 'get by'... Every teacher who wants to practice formative assessment must reconstruct the teaching contracts so as to counteract the habits acquired by his pupils"
>
> (*Perrenoud, 1991 talking of pupils in Switzerland*)

The ultimate user of assessment information which is elicited in order to improve learning is the pupil. Here there are two aspects—one negative, one positive. The negative is illustrated by the above quotation. Where the classroom culture focuses on rewards, 'gold stars', grades or place-in-the-class ranking, then pupils look for the ways to obtain the best marks rather than at the needs of their learning which these marks ought to reflect. One reported consequence is that where they have any choice, pupils avoid difficult tasks. They also spend time and energy looking for clues to the 'right answer'. Many are reluctant to ask questions out of fear of failure. Pupils who encounter difficulties and poor results are led to believe that they lack ability, and this belief leads them to attribute their difficulties to a defect in themselves about which they cannot do a great deal. So they 'retire hurt', avoid investing effort in learning which could only lead to disappointment, and try to build up their self-esteem in other ways. Whilst the high-achievers can do well in such a culture, the overall result is to enhance the frequency and the extent of under-achievement.

The positive aspect is that such outcomes are not inevitable. What is needed is a culture of success, backed by a belief that all can achieve. Formative assessment can be a powerful weapon here if it is communicated in the right way. Whilst it can help all pupils, it gives particularly good results with low achievers where it concentrates on specific problems with their work, and gives them both a clear understanding of what is wrong and achievable targets for putting it right. Pupils can accept and work with such messages, provided that they are not clouded by overtones about ability, competition and comparison with others. In summary, the message can be stated as follows:

- **Feedback to any pupil should be about the particular qualities of his or her work, with advice on what he or she can do to improve, and should avoid comparisons with other pupils.**

Self-assessment by pupils.

However, there is a further dimension. Many of the successful innovations have developed self- and peer-assessment by pupils as ways of enhancing formative assessment, and such work has achieved some success with pupils from age five upwards. This link of formative assessment to self-assessment is not an accident—it is indeed inevitable.

To explain this, it should first be noted that the main problem that those developing self-assessment encounter is not the problem of reliability and trustworthiness: it is found that