Prospects (2014) 44:487–501
DOI 10.1007/s11125-014-9329-7

VIEWPOINTS/CONTROVERSIES

Assessment and the aims of the curriculum: An explorer's journey

Paul Black

Published online: 19 November 2014

Abstract This article considers lessons learnt through involvement in several assessment projects. Early experience, in university work and in school examinations, led to an opportunity to help establish a novel system of assessment for an innovative school curriculum. Different lessons were then learnt from work on a national survey of school students' learning of science, and different lessons again while leading a group to advise the UK government on a new scheme for national testing of all students. Many welcomed the group's advice but politicians rejected it; however, the recoil from this defeat led to very rewarding work on formative assessment. The article ends with reflection on the conflict between the summative and the formative and ways to resolve that conflict, along with the full benefit of formative approaches that investment can secure to help teachers share responsibility for high-stakes summative assessments.

Keywords Assessment · Curriculum · Classroom · Innovation · Learning · Education reform · Formative · Summative

Assessment is a contentious feature of education. On the one hand, the need for evidence, to fairly evaluate individual students and to guide reforms of national policies, is indisputable. On the other hand, the pressures of accountability, with its negative pressures of teaching to the test, lead many teachers and their students to wish that it would all go away.

Given that such tensions infest all aspects of teaching and learning, one might expect that any theory of pedagogy would include some way of clarifying the role of assessment. Yet, in fact, assessment has received scant attention in the literature on pedagogy. Most authors use pedagogy as an inclusive term to cover all aspects of teaching and learning. For example, Alexander (2008, p. 47) defines it as "the act of teaching together with its

P. Black (✉)
Department of Education & Professional Studies, King's College London, Waterloo Bridge Wing, Franklin-Wilkins Building, Stamford Street, London SE1 9NH, UK
e-mail: paul.black@kcl.ac.uk

attendant discourse of educational theories, values, evidence and justifications". He continues: "It is what one needs to know, and the skills one needs to command, in order to make and justify the many different kinds of decision of which teaching is constituted. Curriculum is just one of its domains, albeit a central one".

He goes on to list the core acts of teaching as task, activity, interaction and assessment, but then gives little attention to the last of these, assessment.

In this article I explore this confused area by way of a personal biography. I see this as an attempt to link, in a simple model, several features of assessment practice. This is a modest attempt to inform debate; it would be foolishly ambitious to claim to provide a new and final theory of pedagogy, in which a concept of assessment would be an integral component.

My journey in this sphere began from a narrow base, and only step by step did I realize its wider implications; thus, it makes sense to present this exploration autobiographically. Such an approach has obvious dangers, but I think that the first-hand and personal focus that I can use for this exploration may justify the risk.

From practice to theory

I had my first experience of assessment when I worked on a team marking the national examinations for physics. These examinations were of students at the most advanced level, who aimed to proceed to university studies involving physics. I was impressed by the care taken to ensure that all the team members marked to a well-defined marking scheme and that they shared a common understanding of the standards these schemes represented. At the same time, as a university faculty member, I had to set examinations on the courses I taught; whilst my questions were checked, I was free to use my own judgment in marking them.

These two activities overlapped and developed. In the school-level marking, I was eventually invited to join the team that set the questions; then I had to learn, from experienced examiners, how to ensure that my questions reflected the aims of the published syllabus both in the questions themselves and in my proposed marking schemes. At the same time, I became more critical about the examinations which I, and my faculty colleagues, set for undergraduates. So I became known amongst those colleagues for my concerns about examinations.

My interest developed further when my head of department asked me to give a lecture on assessment to a local meeting of the Institute of Physics (IoP), the professional association of physicists in the United Kingdom. The national IoP had asked its several regional groups to set up local talks and discussions, and in my region that group turned to the university physics department to provide a speaker.

To make it a worthwhile lecture, I spent several days in the library reading all I could find about examinations. This gave me an understanding of issues which I had only superficially grasped at my practitioner level. Two lessons stood out. One was about reliability, and the various factors which threaten the reliability of the results of an examination. The other was about validity, and the relationship between validity and reliability.

My talk was well received; I was invited to publish a paper in the Institute's bulletin (Black 1963), and then invited to join the group whose work had provoked the IoP region to ask for a talk. This group was collecting the test papers, and the other sources of evidence, that twelve different university physics departments were using to determine the

final degree classifications of their students. The group also collected the marking schemes used in assessing the papers. In order to compare these varied and diverse sets of evidence, we classified the demands of each question in the sets of papers according to what it required:

(i) factual recall of material in standard textbooks;
(ii) applications, including application of standard material to a problem, or reorganization of such material for a particular purpose; or
(iii) a mix, with part of the question of type (i) and part of type (ii).

For the courses we examined in detail, the mean percentages of marks assigned to these three categories were 46%, 39%, and 15%, respectively. What was more surprising were the extreme values: one course assigned 66% to factual recall while another assigned only 26% (Black 1968).

Later, I saw these findings in another way when I became an external examiner for several universities, routinely scrutinizing examination results and the decisions based on them, for every department. The external examiner is a senior physics academic from another university who is asked to approve proposed test papers, audit the final marking, and interview a sample of students. In this role I had to convince the university that the degree results were of a satisfactory standard, comparable with those in other universities. In trying to meet this responsibility, I saw evidence of a lack of expert care in some universities' examining systems, and a lack of awareness of the abilities they were actually testing. Overall, having learnt about the theories and practices of summative assessment, I was aware of the very uneven awareness and application of assessment expertise across summative assessment systems.

Inventing summative

My university appointed four staff members to join the governing board of the Joint Matriculation Board, one of the country's school examination authorities. As one of the four, I took part in discussions about both changing the syllabus in physics and strategies for composing examination papers. This experience equipped me well for a surprise invitation.

An independent UK body, the Nuffield Foundation, was promoting the development of new curricula for school science, including a radical revision of the school course at the advanced level in physics, where it could be one of three subjects students aged 16 to 18 chose to study. A team appointed for this work had made unsatisfactory progress and had to resign. Amongst their troubles were problems about introducing simplified versions of topics hitherto considered too advanced at this level, so the foundation set up a new team of five experienced school teachers and one university physicist. I had had no contact with the earlier work, so the invitation to take on the physicist role, and to share the direction of the project with one of the other five, was a challenge. I accepted, and for two years I had a half-time contract which released me from my university teaching.

The project was a year behind schedule and there was great urgency. The other members of the team taught me a great deal about teaching, in particular that my comfort with my delivery of clear and well-organised lectures was misplaced. Clear delivery might be a necessary condition for helping learning, but it was not a sufficient one.

Two features of this experience stand out. First, we aimed to introduce, at school level, ways to study advanced topics, notably, quantum theory and the 2nd Law of

Thermodynamics. The existing school syllabi contained no treatment of any physics developed after 1900—thereby ignoring some revolutionary advances—and these needed to be offered in ways that were both authentic and accessible to the students.

Projects and texts from other countries, my own experiences of teaching undergraduates, and the genius of my co-director, Jon Ogborn, all helped to achieve this aim (Black and Ogborn 1972). I realised that curriculum development could involve far more than reshuffling the well-known topics. However, the team was only able to design radical change because it had 5.5 full-time members, and two to three years, to compose and evaluate the novel approaches. Given how quickly science advances, this kind of resource must be made available periodically; otherwise, students may be taught theories and methods which have been transcended, even contradicted, in the world of practice.

The second outstanding feature was establishing the end-of-course examination for this new syllabus. As I had the most experience in such work, this became my particular responsibility. For all such courses, the final examinations were expected to require two 3-hour test papers set and marked externally, together with assessment of practical work by the students' own teachers. The task was to compose, within this framework, methods which would satisfy one of the national boards that were monitoring and administering the examination, so that it met the standards expected. At the same time, we had to talk with representatives of university departments who would be selecting their students on the basis of their examination results.

Because of the compressed time scale, the team members were inventing the examining approach at the same time as the course was being developed and the teaching materials were being composed for trials in selected schools. The outstanding priority was that the examination should reflect the aims of the course: for example, since an emphasis on learning formulae by heart might conflict with the aim of understanding physics, each student was given a sheet of about 75 formulae and relationships to use during both the course and in the examination.

Because all methods suffer from their particular systematic errors, we judged that an examination which used a variety of methods would be less sensitive to the weaknesses of any one of them. We also felt that parts of the examination should offer no choice of questions; this would make it possible to vary the style and difficulty of questions whilst ensuring that all candidates were assessed over the same variety of levels of demand. These considerations led to a structure using six instruments, as outlined in Table 1. Of these, two were in each of the written paper sessions and two were based on teachers' assessments.

For the coded answer questions, a typical question might present a formula for a phenomenon not treated in the course and test the student's abilities to manipulate it and to relate it to a possible graph; another might reflect the aim of understanding the nature of enquiry in physics. Such questions would assess the student's understanding of physics and of methods required in physics.

A typical question set in the "short answers paper" would explain a situation and ask three or four detailed questions calling for predictions, explanations, numerical calculations, or interpretation of curves on a graph. Some situations might be from the course, others quite new.

These two papers tested many different topics in the course, so that the other written papers did not have to cover the syllabus comprehensively. The 2-hour "long answers paper" offered six questions from which three had to be answered. One type of question would set out a point of view about a general issue at some length; the student would be asked to discuss it, making reference to particular examples. A second type might present a collection of data and ask students to select and combine these to construct quantitative

Table 1 Structure of the Nuffield A-level physics examination

Examination session	Title of paper	Time (min.)	Structure of paper	Weight in assessment (%)
First	Coded answers	75	40q no choice	21
	Short answers	90	8q no choice	21
Second	Long answers	120	3q choose from 6	21
	Comprehension	60	6q no choice	10.5
Third	Practical Problems	90	8q no choice	16
Teacher-assessed investigations, taking the equivalent of two weeks of A-level classroom time				10.5

Source: Black & Ogborn (1977), p. 12

arguments that bore on the problem. The principles involved would be in the course but the particular topics might be new. Several years later, this paper was replaced by a project essay, taking about two weeks of normal classroom time, on a topic the student chose, providing a more valid opportunity for research into sources and for reflective thinking and composition (Morland 1994).

The other paper in this session presented an account of an application of physics which was not in the course, and asked students to answer several questions designed either to test their understanding of the passage and of relevant underlying principles, or to have them evaluate the arguments offered. One aim of this paper was to encourage students to read about physics and its applications beyond the set course topics.

Assessment by the students' classroom teachers involved two different components. Taken together, they reflected the aims of understanding the nature of physical inquiry and of learning to enquire. For the first, a 90-minute practical problems test, each teacher had to set up the prescribed apparatus for eight test items; then eight students could circulate, spending eleven minutes on each. Some might call for use of standard equipment to make measurements, others to make inferences from data, and others to observe a novel phenomenon and suggest how they would plan to investigate it. For the second component, each student had to carry out his or her own investigation, taking approximately two weeks of course time, and write an account of their strategy and findings. Schools were provided with a list, eventually of about 250 topics, so that they could give a different one to each student. Teachers themselves assessed these, but had to submit specified samples of papers, comments, and marks for external checking.

This collection was a unique departure from the conventional examinations. The variety of types and styles of assessment ensured validity because its various types of demands reflected the full range of course aims, and could also reveal students' differing strengths and weaknesses. I described it to a committee of the Board on Testing and Assessment of the US National Academy of Sciences as an example of a comprehensive and balanced assessment system (Pellegrino, Chudowsky, and Glaser 2001, pp. 253–255). The committee's members admired its qualities, but said clearly that it would be far too expensive to implement such a scheme in their country.

It is demanding work to formulate a large-scale examination system which relies on instruments which have to be externally set and marked. Those formulating the tasks should be fully versed in, and share, the aims of the course they are examining. They should also have expertise in composing several different types of question. Apart from the

coded answer paper, all other tests also need trained markers, and expert checking of their work. For the students, each of the six types of task set out in Table 1 calls for its own particular skills; in my experience students only build up such skills slowly through the care with which the course is taught and assessed by their teachers.

This experience showed me the flaws in the inexpensive products commonly used in our public examinations: they undermine the exams' validity, bias the results in favour of students who can focus on their narrow range of demands, and constrain teachers to teach within this range. The reluctance to invest the resources needed to achieve excellence in both curriculum and assessment leads to outcomes which are unfair to students, and which mislead those who interpret the results as evidence of a sound understanding of the subject.

Freedom to roam

In 1978, the government's Assessment of Performance Unit (APU) awarded a contract to the science education groups at Chelsea College and the University of Leeds to carry out national surveys of how school students performed in science. Our brief was to carry out surveys at three ages—11, 13 and 15—with samples of about 10,000 students at each age.

We copied the strategy of matrix light sampling from the National Assessment of Educational Progress (NAEP) surveys in the United States. This strategy starts from the fact that for any one sub-test, one might only need a sample of about 300 to 500 students to obtain an adequately reliable picture of national performance. So the best way to use the committed sample of 10,000 was to distribute each of 35 different sub-tests to 35 different sub-groups selected from the 10,000. The schools that took part were selected to represent the national distribution of schools across such features as school catchment areas, region of the country, and gender balance. Each school might be asked to subdivide its students into two or more groups, with each group attempting a different sub-test. It would not be possible to produce a reliable report on the achievement of any individual school, but each test would produce results which would reflect the overall national performance.

The monitoring at age 13 conducted in 1983 used 35 packages including 465 different questions. For a student to take all of the practical and written tests in this set would have involved about 35 hours of work. It was evident from the results that if all students had to take the same test, then such a test would be limited to only one or two questions for any one criterion, and that could not give a reliable result, even for the average of a large group, let alone for one individual. Thus, we had a unique opportunity to escape from, and then expose, the limitations of any single national test.

However, to make the best use of this opportunity, we had to specify the qualities of school science performance that we should try to assess. We were left quite free to make the choices involved, and produced a scheme of six main categories of performance, as set out in Table 2; the subcategories emerged from our many struggles to frame questions which could distinguish unambiguously between the several dimensions within each category. What is clear from this list is that the approach gave priority to the "process" aspects of science, with only one category devoted to assessing conceptual understanding. The investigations (categories 5 and 6) assessed the combination of the processes (categories 1, 2, and 3) with the concepts (category 4). These choices reflected current beliefs about the desirable aims of science education: one of the team's main tasks was to exemplify these aims in terms of concrete activities students could engage in. The demands of the task sharpened the meaning of each aim.

Table 2 The six-category scheme of APU Science

1. USE OF GRAPHIC AND SYMBOLIC REPRESENTATIONS
- reading from graphs, tables, charts
- representing data as graphs, tables, charts
- describing the relationships that are illustrated by graphs, tables and charts
2. USE OF APPARATUS AND MEASURING INSTRUMENTS
- using instruments
- estimating
- following instructions
3. OBSERVATION
- making and interpreting
4. INTERPRETATION AND APPLICATION
- interpreting presented information
- applying concepts from biology, chemistry and physics
5. PLANNING INVESTIGATIONS
- planning parts of investigations
- planning entire investigations
6. PERFORMING INVESTIGATIONS
- performing entire investigations

As soon as we were in a position to publish examples of the test tasks, with data on the range of students' responses, the attitudes of science teachers to the APU exercise were transformed, from suspicion of the government's intentions, to interest in the test questions. They could see that these were valuable teaching materials—which they could not invent themselves, without the time and experience the team had available. Publishing these questions in a series of eleven short booklets for teachers became an important part of the team's activities.

A different outcome emerged when adult groups, including science teachers, were shown some items and asked to predict the average success rate for students at age 11. There was no clear consensus: a typical result for one question was an average prediction of 70% success, whereas the students' average was 93%; on another question they predicted about 45% and the students averaged 28% (Black, Harlen, and Orgee 1984). In this exploration of students' skills in science process, we were in uncharted territory.

The APU work established many detailed lessons about assessment; I present six of them very briefly here. For more details, with examples, please see Black (1990).

1. To assess students' skill at expressing patterns in numerical data, the data were presented as two columns with the data in the numerical order of a direct relationship, or with the data in the order of an inverse relationship, or in random order; students' success rates differed widely amongst these three. A further variable was the quality of the answers: some presented a straightforward statement of the relationship, while others merely commented on one extreme example and still others attempted to explain the relationship, even though the instruction was only "Describe the relationship". How should we report on the overall ability to "see patterns in presented data"?
2. For questions in category 2, many students would show that they could use instruments correctly, but then, in a category 6 question, they would fail to use this skill.

3. In category 6, some students' choices of strategy showed that they had failed to conceptualise a problem in a way that would allow them to investigate it systematically. This would happen even when they were given a collection of the necessary equipment alongside the question. Their process skills were useless if they could not identify the relevant variables in their model of the task.
4. The same problem was presented to some students in scientific language with a choice of laboratory equipment, and to a matched sample in everyday language with kitchen equipment adequate to the task. The performances on such criteria as accurate timing using a clock could differ widely, a typical result being 54% success in the scientific context and 26% in the everyday context.
5. Many similar variations were linked to the way a task was presented: one task was presented to some students in prose form, and to a second group as prose illustrated by a picture of relevant equipment. In both cases they were asked to describe the investigation they would perform. Those in a third sample were given the actual equipment and asked to carry out the investigation. The results were very different. For example, on the criterion of using "an accurate quantitative method", the success rates were 12% for the prose presentation, 24% for the pictorial presentation, and 43% for the actual investigation.
6. For category 3, observation skills, the lessons learnt in interpreting the wide variations in the results of the teams' first tasks finally led to the following conclusion, for the work with age 15 students:

> Thus, while making and interpreting observations is included for testing in the A.P.U. science framework of scientific activity categories, it may well be that the appropriate place for its specific inclusion in taught science is a practical test closely related to the students' conceptual knowledge base. (Black 1990, p. 25)

The idea of a concept-free "process skill" in investigation was no longer tenable.

The APU conducted four national surveys of science in the years 1980 to 1983; several limited explorations of specific features followed, but the APU was closed down in 1989. It did lead to other projects: one at Leeds explored students' learning of science concepts (Driver, Guesne, and Tiberghien 1985) and one at King's College (Black, Fairbrother, Jones, Simon, and Watson 1992) explored open-ended investigations in science.

Three main lessons about large-scale testing emerged from this process. First, tests must use a wide range of methods: the narrow range of conventional testing gives an invalid picture of students' capabilities and fails to illuminate important lessons about their learning. Second, a short test cannot give a valid guide to the learning that students may have achieved: its sample, of topics and of skills, is too small. Third, the lessons learnt from this process should be of serious concern to all involved in national high-stakes assessment systems.

On this last lesson, my statement from a quarter century ago still bears repeating:

> 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. However, in order to do this, teachers need help from substantial programmes aimed to support teacher assessment with resources of questions, procedures and in-service training. (Black 1990, p. 25)

That quotation can serve as an introduction to my next section, and indeed to much of my subsequent work on assessment. I had become convinced that present systems were seriously inadequate, because they were seriously invalid, and thereby unjust to students, and additionally harmful in distorting the teaching which had to prepare students for them.

Hope frustrated

In 1987 I was invited to chair a UK government task force, entitled the Task Group on Assessment and Testing (TGAT), to advise on the institution of national testing. This was to involve annual assessment of all students at ages 7, 11 and 14, with the results published for every school. This invitation involved both dangers and opportunities. For my present purpose, I need only to quote three of the final recommendations (DES 1988a, para. 227, section XXIII):

> 11. Teachers' ratings of pupil performance should be used as a fundamental element of the national assessment system. Just as with the national tests or tasks, teachers' own ratings should be derived from a variety of methods of evoking and assessing pupils' responses.
>
> 14. The national assessment system should be based on a combination of moderated teachers' ratings and standardized assessment tasks.
>
> 15. Group moderation should be an integral part of moderated teachers' ratings and the results of national tests.

These three arose from some basic principles spelt out near the beginning of the group's report, as the following extracts illustrate:

> - the results should provide a basis for decisions about pupils' further learning needs: they should be formative … (paragraph 5)
>
> … no system has yet been constructed that meets all the criteria of progression, moderation, formative and criterion based assessment set out in paragraph 5 above. (paragraph 13)

The report also included recommendations about the nature of the externally set tests. For example:

> 7. The national system should employ tests for which a wide range of modes of presentation, operation and response should be used so that each may be valid in relation to the attainment targets assessed. (para. 227, section XXIII).

The group's report was published in January 1988, and a collection of supplementary reports dealing with details of implementation, particularly of the organisation of inter-school moderation, was published about two months later (DES 1988b).

The following extract from the memoirs of the prime minister, Margaret Thatcher, explains the subsequent story:

> Ken Baker [then minister of education] warmly welcomed the report. Whether he had read it properly I do not know: if he had it says much for his stamina. Certainly I had no opportunity to do so before agreeing to its publication…. [T]hat it was then welcomed by the Labour party, the National Union of Teachers and the Times Educational Supplement was enough to confirm for me that its approach was suspect. (Thatcher 1993, pp. 594–5)

Baker, who expressed support for the TGAT proposals, was soon replaced. The proposed system for inter-school moderation was dismissed as too cumbersome. Then it was decided that the results of external tests should be published separately from the results of teachers' own assessments. Inevitably, the test results were the only data to which the press and ministers paid any attention. Work to broaden the range of methods used in order to secure the validity of the national assessments was abandoned after a few years of trial. The new minister of education, Charles Clarke, dismissed the novel methods as "elaborate nonsense", and, despite the evidence that teachers had come to value them, the contract to develop them was cancelled and they were replaced by conventional written tests. In a subsequent lecture, Clarke mentioned "the British pedagogue's hostility to written examinations of any kind", which he declared was "taken to ludicrous extremes" (Lawton 1994).

The widespread acceptance of the TGAT recommendations, with ministerial support from Ken Baker, was a challenge to the politicians in power, and it took a few years to undermine its effects and to implement the plan for national testing which they had intended all along (Black 1997). Their firm beliefs, which combined a folk memory of their own school experiences, complete ignorance of such issues as the validity and reliability of assessments, and suspicions of the left-wing motives of "academics", meant that they dismissed any critique of their commitment to formal written tests as the way to raise standards.

Assessment in classrooms

From my work on both the APU and TGAT, I saw the need to further study assessment by teachers. I first reviewed the science education literature in an article about the assessment work done by science teachers (Black 1993). Then a group of researchers, the Assessment Reform Group (ARG), invited me to undertake a broader review of research studies about teachers' formative assessment.

The ARG was a group of British academics, which originally formed within the British Education Research Association, and later operated independently. Having recently retired, I was happy to take on this task, and invited my colleague, Dylan Wiliam, to work with me because I valued the experience and critical insights I knew he would bring to the task. The review, Black and Wiliam (1998a), has been widely cited. We could now claim that a wide range of research evidence supported our belief that formative assessment by teachers would improve students' learning. To publicise this finding, we wrote a short booklet entitled *Inside the Black Box* (Black and Wiliam 1998b); over 55,000 copies have since been sold.

But we had to do more, because we believed that research findings could not lead to classroom change unless they were translated, through closely supported trials, into teachers' classroom practices. We obtained a grant for a research and development project and recruited teachers of mathematics and science (initially) and of English (later on) in each of six schools. The project lasted two and a half years. All participants met at five-week intervals; two research staff also visited the schools, observed lessons, and learnt from, and gave feedback to, individual teachers.

We explained to the teachers that we were presenting them with research findings which had been shown to improve learning. We were not claiming that these were recipes for success, but said that they would be transforming the findings into practical working knowledge, not merely "applying" them (Black and Wiliam 2003). At first, many of them

were puzzled by this approach, but, as the project proceeded, they gradually took over more of the agenda with their reports to, and exchanges with, one another about their experiences. To add to the data collected from the meetings and the researchers' visits, we asked the teachers, at the end of the project, to write their own reflections on the impact the project had on their work.

We published our analyses of the findings in a booklet and a book (Black, Harrison, Lee, Marshall, and Wiliam 2003a, 2003b). In both, we included quotations from the teachers. These publications have proved extremely popular; the researchers and several of the teachers were repeatedly invited to speak about the work. We had also collected data from within the schools which produced evidence that the work did improve the learning of the students in the classes involved (Wiliam, Lee, Harrison, and Black 2004). Government initiatives in the United Kingdom have disseminated the findings in various ways; by far the most effective was the strategy of the Scottish Ministry of Education, which successfully replicated the project's strategy with selected schools, with help from the researchers and teachers involved in the original project (Hallam, Kirton, Peffers, Robertson, and Stobart 2004).

Why was this such a successful project? One reason was that it started with a strong research base. A second was its basic assumption that we could use the research findings to generate new practical working knowledge. Thus, the project's publications contain many insights and findings of which one could hardly find a trace in the original research review.

Since we first publicized our findings, my view of their significance has changed. The original emphasis was a pragmatic one, although the King's College team was relating the work to theoretical analyses based on theories of learning. The subsequent developments can be explained in relation to the four main areas of classroom activity which formed the framework for the initial findings.

The first of these areas was entitled "Questioning": open questions create opportunities for students to express their ideas, giving teachers feedback to guide a formative response. However, there is a vast difference between the response which corrects the student's "error" and one that asks that student to explain why he or she made that response, or invites other students to give their responses to support or contrast with the first reply. The second kind of response can draw the class into a dialogue where they can explore several ideas. Alexander (2006) emphasizes the importance of such activity: "Children, we now know, need to talk, and to experience a rich diet of spoken language, in order to think and to learn. Reading, writing and number may be acknowledged curriculum 'basics', but talk is arguably the true foundation of learning" (p. 9).

Thus, to open up a dialogue, with and between students, is not merely a tactical choice; it could contribute significantly to their development as effective learners. Wood (1998) emphasizes this point in a different way:

> Vygotsky… argues that such external and social activities are gradually internalized by the child as he comes to regulate his own internal activity. Such encounters are the source of experiences which eventually create the "inner dialogues" that form the process of mental self-regulation. Viewed in this way, learning is taking place on at least two levels: the child is learning about the task, developing "local expertise"; and he is also learning how to structure his own learning and reasoning. (p. 98)

Further perspectives on this issue are explored in the extensive literature on classroom dialogue and on self-reflection. However, a teacher who wishes to encourage classroom dialogue has to tackle two problems. The first is to establish routines and expectations that encourage all students to contribute. The second is to achieve a balance between closing

down dialogue before it reaches its full potential, and letting it diverge so widely that the original purpose is lost. No general rules apply here; the teacher's optimum degree of "steering" depends on many features of the overall context. Details about these and other issues I review below can be found in Black and Wiliam (2009), where we discuss our theory of formative assessment.

The second of our four areas focused on teachers' feedback on written work. We started from evidence that feedback that included marks or grades, with or without comments to guide improvement, had almost no positive effect on learning, whereas feedback given only as comments did produce significant gains (Butler 1988). We found two reasons for this difference: marks were a judgment, not an aid to learning, and students ignored comments where a mark was provided. Teachers had two difficulties in responding to this evidence. First, their school policies, and the expectations of students and their parents, usually demanded that they give marks. Second, if feedback through comments was to be formative—to help improve each student's learning—it would have to be carefully aligned to meet the different needs of each student. Such feedback should be seen as another form of learning dialogue, and as such should lead each student to respond to the comments by correcting or rewriting their work.

Butler's research and the more comprehensive studies of Dweck (2000) have shown that different forms of feedback have profound effects on students' view of themselves as learners. When the emphasis is on marks, students develop the view that they are smart or dumb, and moreover that these are fixed, innate qualities. Then high achievers may become reluctant to take risks and thus may not adapt to changes, between schools and beyond schooling. Low achievers will come to believe that there is no point in trying. But feedback offered through comments has a different effect: It produces a growth mind-set in which all students are led to believe that they can improve by their own efforts whether or not they have succeeded so far. Dweck's studies have shown that these different forms of feedback in school can affect students' capability to cope with the challenges in their adult lives.

The third main type of development with teachers was to place new emphasis on students using peer- and self-assessment. This could be particularly useful if students could be engaged in scrutinizing their own and one another's work, whether homework, or attempts at summative tests, by appraising it in group discussions. Seeing their work assessed by their peers, and also seeing how others had tackled the same tasks, helped them to reflect on the ways they had thought through the tasks (recall the quotation above from Wood). Furthermore, students could only evaluate judgements in the light of the relevant criteria for the quality of the learning outcome. Such criteria can only serve as guidance if students understand their meaning, and they can develop such understanding through discussions aimed at applying them in relation to concrete examples. These two features, of self-assessment through peer-assessment, and of realizing the need to clarify criteria to guide one's learning, are key elements in developing meta-cognition, and thus becoming a more confident and independent learner.

This third main area led naturally to our fourth: the formative use of summative tests. If students were given time, after the teacher finished teaching a topic, to engage in feedback on test performance, and to do further work on any difficulties that it exposed, then teachers found that students came to see the test as a valuable part of the learning process. This makes it clear that the distinction between formative and summative lies in the purpose for which the assessment findings are used. This is not to deny the importance of the summative function, but it helps to change the perspective of conflict in which many summative systems are seen as inimical to good learning.

In a recent project (Black 2013), I explored the teachers' skills and practices of summative assessment. In the process, I developed a model of the role assessment plays in pedagogy; it locates assessment within the five steps involved in planning and implementing any piece of teaching and learning, and I believe it can be a useful guide:

A. First, clarify the aims; this often involves a balance between different priorities.
B. Plan the classroom activities which might best secure these aims.
C. Implement them in the classroom, through formative interactions.
D. Engage in an informal summative assessment designed to show up any weaknesses which will need attention if they will undermine future learning.
E. Engage in a formal summative assessment to give all stakeholders guidance to inform decisions about further choices to be made by or for each student.

There are many interactions between these five steps; they cannot imply a single linear sequence. Results from D or E can lead back to further work in B and C in the short term, but in the longer term they might provoke those involved to reconsider the meaning and implications of the aims in A. Overall, teachers could construct positive interactions between all five steps for those school years when they were in full control of the formal summative work; this finding has emerged in several countries. However, for those years when narrow tests, linked to accountability, made E an external imposition, conflict was inevitable.

Thus, whilst the formative approach to teaching can make very positive contributions to a fundamental aim of school education—to build up the capabilities of students as confident and independent learners—the pressures of external testing undermine this approach.

Assessment in pedagogy: Conflict or enrichment?

There is nothing wrong with "teaching to the test" if it is a valid test. A test is valid when its users can justifiably take good scores on the test to mean that the learner is fully competent in understanding and using that subject's achievements. However, inexpensive written tests do not, and cannot, meet this requirement.

All of the lessons I have outlined above reinforce the view that validity in summative assessments requires two revolutions. One is in developing the skills and procedures used in the year-on-year summative assessments by teachers and their schools, as recommended in the TGAT report. The other is to convince those who have power to determine the methods used for large-scale assessments that their reliance on short formal tests is based on ignorance and is deeply harmful to students. This view has been taken further and reinforced by Stanley, MacCann, Gardner, Reynolds, and Wild (2009):

> [T]he teacher is increasingly being seen as the primary assessor in the most important aspects of assessment. The broadening of assessment is based on a view that there are aspects of learning that are important but cannot be adequately assessed by formal external tests. These aspects require human judgment to integrate the many elements of performance behaviours that are required in dealing with authentic assessment tasks. (p. 31)

The required revolutions can be achieved, and have been achieved, in some state systems (Black 2013). A key component of such developments is that teachers have to share responsibility for the high-stakes assessment of their own students, and that their own confidence in their judgments, and the confidence of all stakeholders, can only be achieved

in a system which incorporates a checking of their judgments through meetings of teachers within and between groups of schools. Experience has also shown that the changes can only be achieved in a design which envisages the need for several years of development, and that teachers involved in such development have found that the new responsibilities that they have for step E have had very positive effects on all aspects of their teaching.

The need is clear and there is enough evidence to show that it can be met, but the lack of insight and commitment among state politicians may be an enduring obstacle.

References

Alexander, R. (2006). *Towards dialogic thinking: Rethinking classroom talk* (4th ed.). York: Dialogos.

Alexander, R. (2008). *Essays on pedagogy*. London: Routledge.

Black, P. J. (1963). Examinations and the teaching of science. *Bulletin of the IPPS, 14*, 202–203.

Black, P. J. (1968). University examinations. *Physics Education, 3*, 93–99.

Black, P. J. (1990). APU science: The past and the future. *School Science Review, 72*(258), 13–28.

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

Black, P. (1997). Whatever happened to TGAT? In C. Cullingford (Ed.), *Assessment vs. evaluation* (pp. 24–50). London: Cassell.

Black, P. (2013). Pedagogy in theory and practice: Formative and summative assessments in classrooms and state systems. In D. Corrigan, R. Gunstone, & A. Jones (Eds.), *Valuing assessment in science education: Pedagogy, curriculum, policy* (pp. 207–229). Amsterdam: Sage.

Black, P., Fairbrother, R., Jones, A., Simon, S., & Watson, R. (1992). *Open work in science: Development of investigations in schools*. Hatfield: Association for Science Education.

Black, P. J., Harlen, W., & Orgee, A. (1984). Standards of performance: Expectations and reality. *Journal of Curriculum Studies, 16*, 94–96.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003a). *Working inside the black box: Assessment for learning in the classroom*. London: GL Assessment.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003b). *Assessment for learning: Putting it into practice*. Buckingham: Open University Press.

Black, P. J., & Ogborn, J. M. (1972). The Nuffield advanced physics project. *Teaching school physics: UNESCO source book* (pp. 354–361). Paris: UNESCO.

Black, P. J., & Ogborn, J. M. (1977). The Nuffield A-level physics examination. *Physics Education, 12*, 12–16.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education, 5*(1), 7–74.

Black, P., & Wiliam, D. (1998b). *Inside the black box: Raising standards through classroom assessment*. London: GL Assessment.

Black, P., & Wiliam, D. (2003). 'In praise of educational research': Formative assessment. *British Educational Research Journal, 29*(5), 623–637.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31.

Butler, R. (1988). Enhancing and undermining intrinsic motivation: The effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology, 58*(1), 1–14.

DES [Department of Education and Science] (1988a). *Task group on assessment and testing: A report*. London: DES and the Welsh Office.

DES (1988b). *Task group on assessment and testing: Three supplementary reports*. London: DES and the Welsh Office.

Driver, R., Guesne, E., & Tiberghien, A. (1985). *Children's ideas in science*. Buckingham: Open University Press.

Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality and development*. Philadelphia: Psychology Press.

Hallam, S., Kirton, A., Peffers, J., Robertson, P., & Stobart, G. (2004). *Evaluation of Project 1 of the assessment is for learning development programme: Support for professional practice in formative assessment. Final report*. Edinburgh: Education Scotland.

Lawton, D. (1994). *The Tory mind on education, 1979–94*. London: Falmer.

Morland, D. (1994). *Physics: Examinations and assessment*. Harlow, UK: Longman.
Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.) (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
Stanley, G., MacCann, R., Gardner, J., Reynolds, L., & Wild, I. (2009). *Review of teacher assessment: What works best and issues for development*. Oxford: Oxford University Centre for Educational Development.
Thatcher, M. (1993). *The Downing Street years*. London: Harper Collins.
Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education, 11*(1), 49–65.
Wood, D. (1998). *How children think and learn*. Oxford: Blackwell.

Paul Black (United Kingdom) worked as a physicist for 20 years before moving to a chair in science education. He has made many contributions, to curriculum development in the Nuffield Curriculum Projects at primary and secondary levels, and to research into learning and assessment. In 1988, he was chair of the UK government's Task Group on Assessment and Testing, which formulated advice on the new national assessment system. He has served as vice president of the International Union of Pure and Applied Physics, on the Research Grants Board of the UK Economic and Social Research Council, on three advisory groups of the US National Research Council, and as visiting professor at Stanford University. He is professor emeritus of education at King's College London, and his work on formative assessment, with Dylan Wiliam and the King's assessment group, has had widespread impact.