KING'S College LONDON

University of London

Assessment of Science Learning 14–19

A report prepared for the Royal Society by

Professor Paul Black
Dr Christine Harrison
Professor Jonathan Osborne
Professor Rick Duschl

March 2004
(published in June 2004)

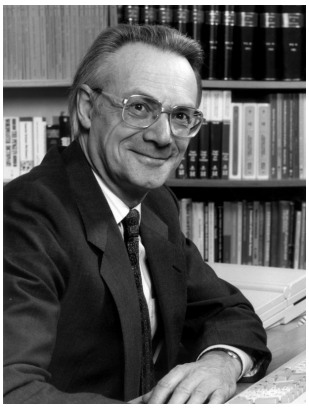
ISBN 0 85403 598 2

Education Section
The Royal Society
6–9 Carlton House Terrace
London SW1Y 5AG
email education@royalsoc.ac.uk

THE ROYAL SOCIETY

Foreword

by Sir Alistair MacFarlane CBE FRS, Chair, Royal Society Education Committee.



In January 2003 a steering group of the Royal Society Education Committee initiated a major research project into the assessment of science learning among 14-19 year old students, undertaken by the Science & Technology Education team at King's College London, and with support from the Gatsby Charitable Foundation. This report is the result of much hard work on the part of the King's College team and the steering group, and I would like to thank them all for their time and enthusiasm, particularly Professor Mick Brown FRS as their Chairman. Their commitment, and the comments we have received from other participants in the study, have convinced the Society that assessment is of critical importance to the future of science education. As the UK academy of science we represent the concern of the science community in the effects that a burdensome examination system and distorted testing regime have on the enjoyment and achievement of young people in science. We are aware that many individuals and organisations share our concerns, and we are optimistic that with their support and co-operation, Government and its agencies will feel strengthened in their resolve to give 21st century schools the assessment system they deserve.

For an electronic version of this report and further information about the Royal Society's education policy work see: **www.royalsoc.ac.uk/education**

For further information about the Science Technology Education Unit at King's College London see: **www.kcl.ac.uk/depsta/education/science.html**

Further copies of this report can be obtained by sending an SAE to:

Education Section
The Royal Society
6–9 Carlton House Terrace
London SW1Y 5AG
email education@royalsoc.ac.uk

Assessment of Science Learning 14–19

Table of Contents:

1 Introduction

In setting out its commission to us, the Royal Society identified four main issues:

(i) providing evidence for the effectiveness of existing dominant models of assessment in science;
(ii) the effects of assessment on the teaching and learning of science;
(iii) a comparison of the ways in which school science is currently assessed against models from other subject disciplines and those from other countries;
(iv) providing recommendations for more effective assessment in school science.

To address these issues, the Society's Education Committee set up a Steering Group to oversee an enquiry, and they invited us to undertake the work.

We have conducted three seminars in which a group has discussed different aspects of these issues and a draft of this report. This group has included three practising science teachers, one headteacher, three local authority science advisers, representatives from the Qualifications and Curriculum Authority, from one of the main examination groups, from the Association of Science Education, from the Institute of Physics, science education experts from the Universities of Leeds, Southampton, York, and King's College, and officers and members from the Royal Society's Steering Group. In addition, others with relevant expertise were invited to contribute to particular seminars. We have also held two open meetings, which attracted colleagues from various sectors of education in London and in Sheffield, respectively, and to which groups of teachers and school students were invited to express their views, and a third meeting with an invited group of parents. Lists of the participants, with reports and details on the papers presented, are given in the Appendices to this report. In the main text, these are represented alongside conventional references to the research literature, as sources that underpin the arguments.

Whilst our work has been carried out in the shadow of the Tomlinson inquiry, our main purpose has been to address the principles and the long-term problems in the assessment of school science rather than to provide a critique of the arguments and tentative recommendations that have appeared in the interim report of that enquiry. Thus, the five main sections that follow here discuss, in turn, the underlying framework for our analysis, then the three main areas of curriculum, teaching and learning, and assessment, and our final conclusions.

2 A Framework

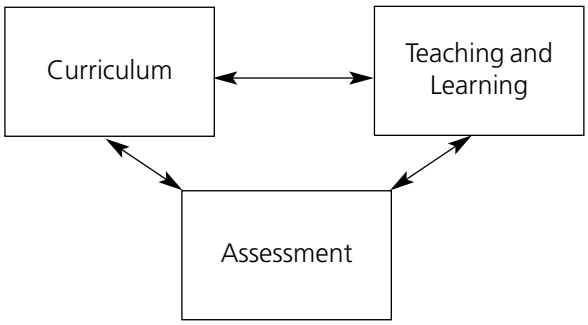
The system of assessment for education over the ages 14 to 19 is being examined in the light of concern over inadequacies and faults of the present system and with a readiness to consider radical change. It may be reasonable to say that things are so bad we must surely be able to do better, but more difficult to say, in positive and proactive terms rather than in negative and reactive ones, what would constitute "better".

The purpose of this report is not to propose the perfect solution: the best hope is for an optimum system, one that meets constraints and resolves conflicts in the best possible way. Our purpose here is to set out and discuss some of the criteria that any re-design of the system should aim to satisfy. The term 'system' is used here to comprise all elements and uses of testing and assessments, both informal in their deployments and effects in the everyday work of schools, and formal, as they are required and implemented by school policies or national policy, and reported and used by students, teachers, parents, the media and policy makers. This view makes clear that there are many stakeholders in this area, and conflicts can arise because each looks to assessment to serve different purposes.

The main types of purpose are concerned, respectively, with the *support of learning*, with *certification*, and with satisfying demands for public *accountability*. The first of these, technically known as formative assessment, is focused on the day-to-day reciprocal feedback in which teachers and students interact in the development of learning. The second is concerned with summative assessment, i.e. with reporting the achievements of individuals; all those who take decisions based on summative results have an interest in this purpose. The third purpose is the particular concern of all the stakeholders, but particularly of policy makers as they strive to promote improvements in the system.

The concern to raise standards over the past 15 years has resulted in numerous initiatives that have affected both the curriculum, and the assessment policies practices and systems that are designed to influence its implementation. But assessment issues are complex, because testing practices are products of our historical and social contexts, and they involve the intersection of aspects of curriculum, psychology, pedagogy, professional competence and status, statistics, budgets and competing public and political priorities. Teachers' views of assessment and of its demands on them can have profound effects on their teaching, even to the extent of making them feel obliged to teach in ways they do not value. Such effects are bound also to affect their interpretation of the aims of the curriculum and may in turn distort their students' experience of those aims. It follows, therefore, that any examination of assessment must be placed in the context of the aims of the curriculum, and of the ways in which it is taught and so experienced by the learners.

Thus, in this report we have adopted a framework in which these three main dimensions, namely curriculum, teaching and learning, and assessment, are each examined in turn. This approach is represented by the following triangle.



This figure represents our view that each of three main elements, curriculum, teaching and learning, and assessment, have to be considered in detail to understand the role and effects of assessment in teaching and learning. However, their mutual interactions are complex and influential. Partly for this reason, the history of initiatives in education is replete with unintended consequences.

In the discussion that follows, we shall, therefore, consider each of these elements in turn. However, because of the strong interactions involved, the discussions must overlap and there may be some inevitable repetition as the same issues are examined from different perspectives. In particular, implications for assessment will be explored in sections entitled Curriculum and Assessment, and on Teaching, Learning and Assessment: these implications will then be taken further in a section on Assessment Issues in which we develop more specific and explicit arguments about the design of assessment systems.

3 Curriculum and Assessment

Any consideration of the science curriculum and its assessment must begin with an understanding of the goals and purposes of science education. Only when these are clearly defined and recognised is it then possible to produce an assessment system that supports and augments the curriculum. The problem for science education is that there are two fundamental aims to teaching science which are essentially in conflict with each other.

Aims in conflict

Traditionally, the science curriculum has been a pre-professional preparation for the next generation of potential scientists. Such an education is, whatever people may say, more akin to a form of pre-professional training for the future scientist rather than an education *about* science and the insights offered by the world-view it presents. The curriculum that serves this aim begins with the foundations of science addressing basic concepts in a piecemeal fashion, which often seem unrelated to the neophyte student. On these foundations are built additional layers of knowledge. Any sense of coherence, based on an understanding of the major themes and interrelationships within the sciences, is only obtained after many years of study. In addition, the knowledge that forms the substance of such a curriculum is well established, unequivocal, and seen as not available for questioning. Consequently, there is a tendency for much of the subject to be taught in an authoritarian manner, leaving little space for discussion or exploration of what, after all, are a set of unnatural and difficult ideas. As Lewis Wolpert has argued, even the idea that day and night are caused by a spinning Earth – an idea that is introduced at Key Stage 2 (KS2) (ages 7–11) – contradicts the evidence of our senses.

One of the unfortunate outcomes of this approach is that students are left with a strong impression that science produces certain and absolute knowledge of a real world. Little space, if any, is provided to explore the uncertainties and tentative nature of the knowledge produced by science-in-the-making, either in the contemporary or historical domain. Few students would be able to recapitulate, for instance, the arguments surrounding the Copernican idea that the Earth moves round the Sun; the experimental evidence that led Torricelli to advance the view that we live at the bottom of a 'sea of air'; or the evidence that led Pasteur to argue that putrefaction is caused by microscopic organisms that pervade the air we breathe. Such knowledge is essential if students are to understand the tremendous intellectual achievement of the scientific endeavour.

Moreover, the dominance of the requirements of the scientific community means that each stage of a science education is always seen as a preparation for the next stage and not an *end in itself*. The approach adopted begins with a set of fundamental but ostensibly unrelated concepts: the parts of the body, the names of the planets, the distinction between elements and mixtures. To begin to approach the edifice of knowledge that rests on such foundations it is then necessary to continue with science post-16 and preferably to university. Such a critique of practice in science education was elegantly articulated by Cohen (1952) and, despite the many structural changes to the English education system, remains largely true today.

In the past 20 years, most societies have insisted that that science should be compulsory for all. The major rationale for this change was a view that the pervasiveness of science within our society required all students to be educated about science until age 16. In short, that a knowledge of science was essential to maximise the engagement in democratic society of our future citizens – an argument aptly articulated by the 1995 European White Paper on Education and Training (European Commission, 1995)

Democracy functions by majority decision on major issues which, because of their complexity, require an increasing amount of background knowledge. For example, environmental and ethical issues cannot be the subject of informed debate unless young people possess certain scientific awareness. At the moment, decisions in this area are all too often based on subjective and emotional criteria, the majority lacking the general knowledge to make an informed choice. Clearly this does not mean turning everyone into a scientific expert, but enabling them to fulfill an enlightened role in making choices which affect their environment and to understand in broad terms the social implications of debates between experts.

Another major argument made for elevating the study of all sciences to the curriculum high table was that permitting girls to relinquish physical sciences and, conversely, boys biology, simply perpetuated unacceptable gender divisions. However, this policy was implemented with little thought given to the nature of the curriculum that would be appropriate to meeting the needs of science for citizens as opposed to science for future scientists. Rather, what happened was essentially a modification and adaptation of the curriculum offered such that the major goals of the curriculum remained unchanged.

The essential tension is that the school science curriculum still has to develop an adequate knowledge base for those who wish to continue to pursue the study of science post-16 while also attempting to develop an understanding about science necessary for citizenship. However, both scholarly analysis and empirical evidence would suggest that the latter aim is best served by a fundamentally different curriculum and approach: one that gives much less significance to content and much more weight to teaching *about* science. The approach to content here is

more top-down rather than bottom-up. From this perspective, the major explanatory themes should be presented in a contemporary context that offers enhanced relevance and an opportunity to explore and appreciate both the intellectual achievement of science and its cultural significance. The curriculum space released by this approach would then permit exploration of how decisions about science and technology affect society, the nature of risk and its assessment, the mechanisms by which scientific ideas are judged and evaluated, and the role of models in science. The core of the new pilot course – *21st Century Science* – is one attempt to offer just such a course.

In summary, since its inception, the compulsory science curriculum has wrestled with the competing demands of attempting to serve the twin masters of the specific requirement that science education should provide a *training* for future scientists on the one hand, and the needs of the majority for a broader scientific *education* on the other. History would suggest that it is the former that has predominated and that these aims are often in conflict, undermining the effectiveness of the curriculum at achieving either.

Moreover, even though science is now compulsory for all, the curriculum has not undergone a significant transformation. In essence, there has been an external change in the structures and agency of the science curriculum without an internal change to match the new aims and goals of science education for all. Old wine has been placed in new bottles and the curriculum for a minority simply adapted and offered to the majority. Yet, science education for all can only be morally justified if it offers something that is of universal value to all.

Recommendation 1
Any framework for assessment must, first and foremost, support a curriculum and pedagogy that meets the needs of all learners.

Building on experience

A product of this tension between the competing demands of any system of science education has been a series of attempts to provide for those who will not continue with science. The first half of the 20th Century saw a succession of general science courses; later, in the 1970s the Nuffield Secondary Science course and mode 3 CSEs were developed, and in the 1980s GCSEs. More recently, in the past ten years, there has been a succession of qualifications falling under the heading of General National Vocational Qualifications (seminar paper: McKay). The latest embodiments of such courses are the applied GCSEs and A-levels. However, the GNVQ has had a troubled history, with successive adaptations and amendments to both the course and its assessment. Moreover, attempts to offer vocationally oriented courses have been bedevilled by the common conception that such courses are of lower status and less worth than academically oriented science courses. Such a lack of stability suggests that no acceptable solution has yet been developed to meeting the needs of students who wish to pursue courses in applied science.

Nevertheless, despite their many troubles, such courses have provided a valuable theatre for non-traditional approaches to assessment which make much more use of teacher assessment and criterion referencing (seminar paper: McKay). Thus, they have provided a means of exploring a different range of course-based assessment, sustaining some of the innovative approaches that were begun through the Mode 3 CSEs and GCSEs and the Graded Assessment in Science Project (seminar papers: M. Brown,).

The most recent, more radical attempt to meet the needs of both constituencies of students is the pilot *21st Century Science* course. This course offers a common core for all students with subsequent, optional academic or applied modules. Furthermore, it also offers several innovative approaches to assessment in science education and to the use of course-based assessment (seminar paper: Nicholson).

In summary, both applied and traditional science education have made significant use of teacher assessment and have offered a variety of models of different approaches to assessment in science education. There is, then, a body of knowledge and experience that can be used to develop appropriate forms of course-based assessment for science education. Thus, there is no need to either reinvent the wheel or begin from scratch.

The effect of assessment on student attitudes and learning

Pupils' views about such a science curriculum have been the subject of two recent studies, both of which have documented considerable dissatisfaction, particularly at GCSE (Cerini, Murray, & Reiss, 2003; Osborne & Collins, 2000). Although certain features such as the opportunity to engage in empirical enquiry do appeal, other elements such as the crammed and content-laden nature of the curriculum, the lack of any opportunities for discussion, the considerable repetition of subject matter, and the overuse of copying have led to the voicing of many well-articulated complaints. In many cases, the outcome has been the formation of negative attitudes towards one or more of the sciences (Gardner, 1975; Osborne, Simon, & Collins, 2003; Schibeci, 1984). Notwithstanding the many exemplary efforts of many enthusiastic and inspiring teachers of science, the recent research would suggest that the pupils' experience is structured by a curriculum that is dominated by content requiring their teachers to 'frogmarch them across the scientific landscape'. The pedagogy commonly used by such a curriculum is heavily reliant on transmission and one-way communication of information, and lacks a variety of teaching approaches.

When I was at school science was about making sense of things and history about facts; these days it's the other way round. Parent

The need to cover *so many ideas* in the short time allotted means that there is no time to stand and stare. For example, to diverge and follow a line of enquiry of interest to the teacher's pupils or to discuss the significance and meaning of a given idea using a diagnostic test item, a concept cartoon, an engaging demonstration or other pedagogic approaches. Rather, many a teacher of science is forced down the relentless path of curriculum coverage, often against their better judgement, simply because this approach is functionally effective ensuring that the teacher meets their moral obligation to their students to cover the entire curriculum (Hacker & Rowe, 1997). However, if such a science education is as effective as some would claim in meeting the needs of society, we must ask why, for instance, are so many adults unable to distinguish a virus from a bacteria, or to know that it takes the Earth one year to rotate around the Sun?

The present system does not meet the needs of any but a small minority of the students I teach. It is based on a specification of content in the National Curriculum that requires students to memorize and repeat facts about scientific knowledge that are of little interest or relevance to them. It does not prepare them to understand the scientific issues they will meet in everyday life. Teacher

Such an approach is undoubtedly effective at generating the kind of superficial knowledge required to recall the information required by standard summative assessments such as KS3 tests or the GCSE exams. However, the lack of any coherence or exploration of scientific ideas too often makes such knowledge seem irrelevant, meaningless and of ephemeral value. The consequence for the majority who abandon all science education post-16 is that they are left with little more than fragments of unrelated knowledge and no appreciation of the cultural achievement of science. In particular, that science offers some of the best and most useful explanations of the material world. In addition, some of these ideas – that life on earth has evolved through a process of natural selection, that energy is conserved, that matter can neither be created nor destroyed, and more – constitute major explanatory themes that frame not only our thinking about scientific phenomena but our Western cultural identity.

The other major criticism of such an approach is that the dominance of factual recall of information leads to an emphasis on the lower levels of Bloom's taxonomy of cognitive skills: that is, factual recall, comprehension and application as opposed to analysis, evaluation and synthesis. The cognitively undemanding nature of such a curriculum then does little to justify the argument that the study of science offers an opportunity to develop the kind of higher-order thinking skills that contemporary societies increasingly demand of their youth. Moreover, for some pupils, particularly the more able, it is an additional reason why school science appears unappealing, leading to the loss of the very students science desperately needs to sustain its own future.

Even the history of the attempt to incorporate and assess students' understanding of the process of scientific enquiry is strewn with good but failed intentions. It has taken the scholarly and professional community of science educators three versions of the national curriculum to develop one that has proved resolute against scholarly criticism (Donnelly et al. 1996; Donnelly & Jenkins, 1999). However, it has not proved possible to develop a valid method of assessment that teachers cannot manipulate to maximise student outcomes (seminar papers: Carson, Parkyn & Wagner; discussion report: Open Meetings).

The quality of the work has also been inhibited by oppressive moderation procedures and by lack of professional training for both teachers and moderators. It is ironic that the only aspect of science that is entrusted, at GCSE level, to teachers' assessment has led to 'Investigations' which the various external pressures have reduced to stereotyped exercises that are widely recognised to be of no interest to students and to present them with a mockery of scientific enquiry. Similar damaging effects of moderation that lead to 'rubric-driven instruction' have been reported in other subjects and in other countries (Paechter, 1995; Baker & O'Neil, 1994).

The consequence is that any school lessons for students exploring or modelling the process of scientific enquiry have become little more than artificial and ritualistic procedures that bear only a distant relation to the process they attempt to emulate. This is not to argue that enquiry conducted in the context of a school science laboratory should be authentic. How could it possibly be? Neither the equipment, techniques nor the nature of the problem have much similarity. For instance, scientists are often working at the boundaries of their knowledge on ill-defined problems. School science, in contrast, works with well-defined problems that exemplify typical challenges. However, what school science should do is attempt to offer authentic educational experiences where, even if the problem is well-defined, the student should be encouraged to work either collaboratively or independently to produce a solution that is a product that reflects their own creative and intellectual endeavour. Moreover, given the significant investment by society in specialised laboratories for the teaching of science, any curriculum should encourage empirical investigation. And, most critically, the mode of assessment should support rather than undermine or negate such practice.

Not surprisingly, then, for all these reasons and more, many students opt out of studying science at the point of

choice, choosing subjects that offer greater room for self-expression, that are taught in a manner requiring more active engagement, and that offer more personal significance, enjoyment or meaning.

As currently constituted, external assessment in school science education would appear to have a malign effect on the teaching of science, encouraging teachers to teach by transmission which, in turn, results in negative student attitudes towards schools science. Too often, assessment in school science supports a practice that sees science as a body of knowledge to be learnt rather than as a way of knowing which has transformed the world in which we live.

Recommendation 2
As well as demonstrating a knowledge of scientific concepts and methods, assessment in school science must require students to demonstrate the ability to analyse, evaluate and reflect on this knowledge and on relevant scientific reports in a range of contexts.

The significance of assessment

For too long the implicit premise of all engaged in curriculum development has been that assessment of the learning goals of a given curriculum is something subsidiary or secondary to selecting a body of suitable and engaging content matched with an appropriate pedagogy. One notable exception is the Nuffield courses of the 1970s. They achieved a better synergy because they had both the time to develop novel forms of assessment and because the assessment system of that era was more responsive and positive towards innovative practice.

However, for most examples of curriculum change, assessment has been essentially an afterthought. Historically, the role and significance of assessment in curriculum development has been undervalued and underinvested. Curriculum development that lacks consideration and development of appropriate forms of assessment is like a cart without a horse: simply unfit for purpose and unlikely to succeed. The mistake here has been a belief that the intentions of the curriculum would be read from the curriculum whereas, in reality, because any terminal examinations are, for both teachers and their pupils, high-stake events, any thoughtful and responsible teacher reads the intentions first and foremost from past items and specimen papers provided by the examination boards. This would not matter if curriculum, pedagogy and assessment were well matched. However, the lesson of history is that the assessment of any curriculum has overwhelmingly been of secondary import, and has had nowhere near the same level of intellectual or financial investment given to curriculum, pedagogy or teacher development. This we believe to be a fundamental error.

Recommendation 3
No curriculum development should be undertaken without simultaneously developing appropriate form of student assessment.

Any science curriculum has four sets of learning goals: conceptual – which is the body of knowledge, understanding and skills to be acquired by the learner; cognitive – which is the development of student's reasoning and thinking skills; epistemic – which seek to offer an understanding of how we know the scientific account to be true; and social – which aim to ensure that learning the subject is enjoyed and to develop other attributes such as students' ability to work collaboratively. The balance between these elements clearly varies between curricula but the intent should be that the assessment of any given curriculum matches its goals. However, a combination of the constraints of summative assessment, which requires simultaneous measuring of large numbers of students in finite time, the minimal use of teacher assessment, and a lack of expertise in the science education community in assessment, has resulted in a significant skewing of these goals. Put simply, it is much easier to write valid and reliable items that assess students' knowledge and understanding of science than it is to measure their cognitive capabilities or their understanding of the epistemic basis of belief. As for social goals, the difficulties associated with their assessment mean that their measurement is neglected by virtually all science curricula. Thus, even if all of these goals are considered important and appropriate to measure, it is, nevertheless, *only the measurable that is important*, particularly in a context of high-stakes assessment. The outcome is that assessment becomes reliant on too narrow a set of performances. Consequently, its reliability is open to question (see the later section on Assessment), as is its validity in that it fails to measure the full range of knowledge, skills and attributes that the curriculum ostensibly aims to achieve.

In the case of science curricula, given their emphasis on foundational concepts, the easiest items to test are those that measure factual recall, comprehension and application. Not surprisingly, it is these items that predominate in tests at both age 14 and age 16. The consequence is that the validity of the examination is at best dubious and, at worst, highly questionable. Particularly when recognising these emphases, teachers of science adapt their pedagogy and curriculum time to maximise the achievement of their students at this limited set of competencies. However, the superficial nature of learning within such an assessment framework is demonstrated by many studies that have examined students' conceptual understanding and found it significantly wanting, not only in a wide range of fundamental scientific concepts (Driver et al. 1994; Wandersee et al. 1994) but also in their understanding

of the nature of science (Driver et al. 1996). This research would suggest that many of the basic ideas of science are either not easily understood, hard, or unnatural and require significant exploration for effective learning.

Second, there is the lack of consensus about learning goals other than those associated with knowledge and understanding. In the case of epistemic learning goals, the science education community and, in turn, the science National Curriculum, has only recently begun to accept that these have a justified place on the curriculum. As a community, there is a virtual absence of any experience and expertise of how to assess student understanding of the relationship between ideas and evidence and their significance in science. For instance, the first attempts to assess this new component of the GCSE curriculum, introduced in 2001, have been subject to many extensive criticisms. An interesting parallel here is the history curriculum. As a school subject, it has transformed itself from a subject dominated by a knowledge of the principal 'facts' associated with a selected set of historical eras to one that attempted to show that history was a process of interpretation, demonstrating that historical accounts are not unequivocal knowledge of the past but more carefully argued accounts based on the best judgements that can be made from limited evidence. The argument of the report *Beyond 2000: Science Education for the Future* was essentially that school science needs to undergo a similar transformation.

Recommendation 4
More research and development is needed by examining groups, teachers and researchers to produce a range of reliable and valid methods of assessment that measure more than the narrow range of learning goals and outcomes of current assessments.

In summary, one way of perceiving the effect of assessment is in terms of the concept of the intended curriculum, the implemented curriculum and the attained curriculum. The first of these is what is defined by the syllabus, the textbooks, support materials and professional development courses. The second is what the teacher actually implements in the classroom, and the last is what is actually learnt by the students. The lesson of history is that in specifying the intended curriculum, there has been an enduring neglect of the significance of assessment to the implemented curriculum. For the intentions of the curriculum are read not from the syllabi but from the method and manner of its assessment. If nothing else, the aim of this report is to make the case for the centrality of assessment to curriculum development. More fundamentally, it seeks to argue that as a society we cannot achieve the science education we seek to offer our young unless we invest our resources in developing models of assessment that reflect our aspirations for the kind of knowledge and experiences we wish school science to offer.

4 Teaching, Learning and Assessment

Principles of learning

Any consideration of effective teaching of science must consider how children learn. Innovations in teaching are increasingly based on results of research into how pupils learn (Bransford et al. 1999), and most approaches now focus on three principles: that learning should start from a learner's existing understanding; that learners must take an active part in their learning; and that learners need to develop meta-cognitive capabilities, i.e. the ability to control and to regulate their own learning. Overall, learners have to construct their knowledge, not merely receive it. Thus through activities such as observing, classifying, experimentation, pattern seeking and comparing and contrasting ideas, students' scientific conceptual understanding can be developed, tested and challenged (Hodson, 1998). Teachers can initiate and facilitate such learning by setting demanding tasks, encouraging open discussion of ideas and providing counter-arguments and challenge.

Some learning experiences result in 'shallow learning' (White, 1992; Entwhistle, 1991), whereby students often exhibit widespread misconceptions and misunderstandings but are able to score well on test items that require recall of terms or manipulation of formulae. Such students have acquired knowledge but, because their knowledge has not received a deeper consideration, it remains superficial and is often discarded and forgotten with time. 'Deep learning', in contrast, involves the student in thinking about the meaning of what they are learning. It is this activity that shapes and strengthens their learning as their understanding becomes part of the personal knowledge of that individual. However, this more productive approach to learning is more demanding of teachers: teachers who 'teach for understanding' rather than those who 'teach to the test' require a much broader range of teaching strategies (Hashweh, 1996).

Nuthall & Alton-Lee (1995) tested students shortly after a course and then again one year later. In the delayed test, students who had learned to respond by recollecting relevant classroom experiences did not perform as well as those who used deductions or inferences to answer the questions. These different approaches were related to the ways in which different groups had been taught, showing that the learning experience determined the long-term effectiveness of the learning. Nuthall and Alton-Lee suggest that there are two parallel systems involved in answering test questions, the first being memory retrieval and the second deductive processing. Developing both capabilities requires a learning experience that models the deductive processes required for the construction of knowledge as well as the retention of facts.

Interactions in the classroom

Teachers develop deep learning in students by interactive feedback with them: they achieve this by finding ways to elicit the students' existing understanding and then helping them to reconstruct and develop their knowledge. This process is a mode of assessment known as 'formative assessment'. There is substantial and rigorous research evidence that where teachers collect rich evidence from the discussion and activities in the classroom and then use it to make professional judgements about the next steps in learning, standards of learners' attainment are improved (Black & Wiliam, 1998). Science teachers may value formative assessment, but are often concerned that time constraints, resulting from a full curriculum and the requirements of extensive summative assessment, prevents them using formative practices to support learning in their classrooms (Daws & Singh, 1996). Nevertheless, important advances have been made in developing this aspect of assessment in classrooms in England, with teachers working under the constraints of our present systems (Black et al. 2002, 2003). Similar success in Scotland has led to such work being a major element in their reform plans for education (seminar paper: Hutchinson).

Other aspects of the interaction between teacher and learner are important because of their effect on the motivation of learners which, in turn, is influenced by their personal beliefs about themselves as learners. Research into these 'self-theories' (Dweck, 2000) shows that they fall into two main categories. One is variously described as performance-oriented or ego-oriented: students in this group believe that people are either intrinsically smart or dumb, i.e. they hold a 'fixed IQ' of intelligence. Such students tend to avoid any challenging task, either believing that the task is bound to be beyond them or that there is a risk of failure which will damage their belief about their IQ. The other category is described as task-oriented or learning oriented. These students believe that they can improve by their own efforts and that they can learn from failures, and are more willing to take on challenging tasks. Students in this second category outperform those in the first and cope better with such changes as the transition from school to university.

Students can be moved between these categories by the ways in which feedback is provided. However, many science teachers create experiences that reinforce the performance-orientated view and inhibit the more productive habits of the learning-orientated view (Hodson, 1998). An example of this would be teachers who regularly give grades, marks or levels on work. Such practices reinforce students tendencies to believe that they are either high ability or low ability, resulting in the former only attempting tasks for which they feel sure they will achieve high grades, and the latter becoming de-motivated because the process is simply regularly

reminding them that they are failures. At its extreme, teachers who believe in fixed intelligence provide, instead of challenging tasks, activities in which learners are encouraged to learn in a rote fashion. In these, students do not have to think much about the work and so learn in a superficial way, the quest being how to get the correct answer rather than to develop understanding. If such a learning environment is also encouraged by the assessment regime, then both teachers and learners become complicit in 'working towards the test'.

Learning science

The effects of assessment on teachers' approaches to teaching and learning are of particular importance in science education, because some of its aims are conceptually challenging. The neglect of 'deep-learning' approaches has led to the well-researched evidence that students hold everyday beliefs about natural processes long after they have nominally learned the more subtle scientific explanations (Driver et al., 1994). In overcoming this problem, it is essential to note that in science lessons, the subject matter that is taught is derived from two main sources: content knowledge (curriculum) and pedagogical-content knowledge (teaching methods, strategies and styles) (Shulman, 1987). Content knowledge is the range, detail and interconnectedness of the basic concepts within the science curriculum. Pedagogical-content knowledge is the knowledge teachers use to present and translate content knowledge such that learners can access the scientific ideas and be supported in developing their conceptual understanding. Tobin and Garnett (1993) stress the importance of teachers having adequate pedagogical and content knowledge so as to assist their students in developing science content from the classroom activities in which they are engaged.

Where, as often happens, teachers put the emphasis on their teaching rather than on their student's learning (Brown, 1998), their focus tends to be on delivery of the content knowledge. While this does not prevent learning taking place (Miller, 1989), it may detract from it as the classroom emphasis can become the coverage of content rather than support of learning. In such environments, assessment can become a means of measuring knowledge transfer where it checks on recall, comprehension and application of knowledge rather than seeking understanding through synthesis, analysis and evaluation (Bloom, 1956). Where the assessment system takes this approach, teachers feel pressurised into selecting activities that engage their students with these approaches (seminar papers: Carson, Parkyn & Wagner).

The assessment system drives teachers to find ways of maximising performance of students which often circumvent the intellectual demands of the subject so that students do not develop cognitive skills through their learning of science. Teacher

Recommendation 5
Priority should be given to the design and promotion of programmes of professional development which help science teachers to teach in ways that develop further aspects of their teaching which are firmly grounded in established principles of effective learning.

Negative pressures

The past two decades have seen many changes in the ways that schools are managed, inspected and financed, with school improvement and the leagues tables becoming key factors in the daily life of schools. Such pressures influence the way that teachers decide to teach, how they manage the curriculum, and the ways in which they interact with their students. The introduction of the National Curriculum in 1989 and its assessment procedures affected teaching strategies in schools (Fairbrother et al.,1995; Hudson & Smith, 1995; Hacker & Rowe, 1997; Russell et al. 1995). All four studies reported a reduction in the range of teaching strategies employed by teachers, with a movement away from pupil-centred teaching towards a more didactic approach. In Hudson and Smith's survey, over 70% of the sample admitted that they could not teach in the way they would prefer because of the curricular and assessment demands of the National Curriculum. They felt they were constrained by the content overload of the curriculum and the introduction of new assessment procedures that had been externally imposed .

Most schools today use examination data to help them make decisions about where to focus their efforts and to check the school's performance from year to year. Although such practices may be beneficial to schools in deciding where to focus resourcing and support, it also affects how teachers interpret their role. The competitive climate in which teachers work can drive them to focus on preparing students for examinations and, although this must be part of what teachers do within their teaching, it can lead to students thinking that examination performance rather than learning is always the goal. An overemphasis on preparation for high-stake examinations not only influences what teachers teach but also how they teach. A recent review (Harlen & Deakin-Crick, 2003) has shown that under the pressures of high-stakes testing, teachers have become expert at teaching to the test, to the detriment of teaching for understanding or for developing creativity, with accompanying negative effects on their students' motivation and enjoyment.

In summary, our emphasis here has been that teachers' everyday methods play a critical role in student learning. Through their actions and practices, they create learning experiences that allow students to engage with the ideas, the methods and the achievements of science. However, the decisions that teachers take in the classroom are governed by their beliefs and moulded by the experiences that they have witnessed both in their own schooling and in their professional lives as teachers, and these in turn are informed

and affected by the culture and systems in which they work and live (Ball, 1981), in particular, by the values implicitly or explicitly communicated by the assessment framework . Their capacity and freedom to work to the best professional standards are bound to be influenced, for good or ill, by the expectations of students, parents, politicians and society.

Recommendation 6
The culture of schools, the curriculum and the assessment systems for science must be reformed to enable and reward good learning practices in classrooms.

5 Assessment Issues

The recommendations from the preceding sections are the starting point for discussion of assessment. The overall message of these recommendations is that any reform of science education must, first and foremost, support a curriculum and pedagogy that meets the needs of the majority. This aim can only be achieved if there is a positive link between curriculum and assessment, and the key to this link lies in its effect on teaching and learning. Any new assessment model should not lead or constrain teachers to teach to the test but rather to teach for understanding. The methods adopted for any new approach to assessment should be designed to achieve as much synergy as possible with practices that promote the learning of students.

Criteria of quality

The next stage in the argument is explore the above issues in the light of the basic criteria of quality that any assessment system should be designed to satisfy. The first of these is **reliability**, which is an indicator of the accuracy of the measure that an assessment purports to supply. If an assessment is to be reliable, then users need assurance that, if a parallel test were to be taken on another occasion, within a short time, the same result would be obtained. This may be an impossible demand, and it would be more realistic to accompany any result with a measure of error, i.e. with an estimate of the probability that the result could differ in grade or mark from the average of many such parallel repetitions. Current testing procedures are comparatively poor in this respect; there is a serious lack of data on the reliability of such tests, but such data as do exist suggest that the chances of candidates in a public examination being wrongly graded could be as high as 30% (see seminar paper: Black). The absence of well-researched data on the reliability of our tests means that they do not meet internationally recognised criteria for acceptance of tests (AERA, 1999), and that users have inadequate information in making decisions based on test results. It also hampers policy decisions on the design on assessments systems, for it is not possible to maximise reliability by an optimum choice between, or combination of, measures from assessments that are obtained by different means, if the reliabilities of these separate measures are not known. This consideration is particularly relevant for any decision that has to weigh the relative advantages of external tests against those of summative assessments by the student's own teachers. Consideration of reliability is also bound to have a wider relevance to assessment design: for example, a system that calls for qualifications that use the aggregation of a wide range of achievement data would be inherently more reliable than a system that provides a large number of qualifications, each based on a narrow range of achievements.

Recommendation 7
Any reformed system of assessment should be grounded in a thoroughly researched basis of evidence about the reliabilities of different sources of assessment information.

The second outstanding criterion is **validity**. Put simply, this involves an evaluation of the extent to which an assessment does what it is designed to do. However, assessments are often designed to do several different things, so validity has become a complex, multi-faceted criterion. It reflects the fidelity between assessment results and the aims of the science curriculum. Current assessments lack validity both because they constrain students' learning in ways that do not reflect authentic aims which that learning should pursue, and because they may mislead users about the qualities which those users expect to see reflected in test results. The weakness seems apparent to parents:

There seems to be very traditional ways of assessing in science and a lot of it and children find it hard. Other subjects have more interesting ways to test and it just seems less stressful for those. Parent

The supplementary note in the appendices gives a brief account of typical science examinations at GCSE and at A-level. For the former, a typical paper calls for about 50 short responses, in spaces assigned on the test paper, only a handful of which allow more than four lines or writing. These very limited and uniform types of question may follow from a view of the examiners that they must sample many different parts of the syllabus within the limited time permitted. They may also be influenced by the fact that short highly specific demands are easier to mark in a uniform and defensible way: fear of variations between markers and of complaints and calls for re-marking may be a factor here. It follows that there is no incentive for candidates to practise writing any extended prose about science, nor to attempt any problem that calls for more than three lines of calculation.

Even more worrying is that the demands in at least some A-levels are very similarly restricted. Written papers are composed of short highly structured questions, and most papers specify a length of only a few lines. One board's system has no course-work component, and uses 90-minute practical tests, again highly structured – a typical practical test calls for 14 separate responses, with only two having more than four lines provided, only three requiring any calculation. Thus there is virtually no opportunity, and therefore no incentive, for pupils to write at any length about the science that they are studying, and no incentive to devise a strategy for tackling a problem as opposed to following pre-ordained steps. For empirical experimental work, the only preparation that seems necessary is to practise exercises that take no more than 15 minutes to complete, for which all the equipment is specified, and for which step-by-step guidance is provided with no more than a few lines of writing required at each step. These restrictions in styles of examination mean that only low-order thinking is assessed; as one parent saw it:

They seem to have a lot of remembering to do in science about lots of different things. I just wonder how much space they have left in their heads to do anything more than regurgitate facts when it comes to exams. Parent

Similar concerns are also frequently expressed, as already pointed out in section 3 above about the coursework assessment of so-called 'investigations' in science. As one teacher expressed it:

My department and myself all have very strong views about the coursework component of GCSE. Unfortunately, because these investigations are mainly not assessing what they are supposed to, we were unable to find anything positive to say about them apart from 'in theory'. Teacher

It is hard to claim, therefore, either that present assessments are valid in reflecting any defensible set of aims for science education, or that they are valid as predictors of ability to succeed in more advanced study in science.

Thus, it is clear that current methods of assessment are far from reflecting and supporting the courses and learning approaches that a renewal of science education will require; they are quite inadequate, as reflections of the range of knowledge and skills that this education should promote. If they are to support the argument for a new approach to science education to meet the needs of the majority at the foundational levels, and do justice to achievements in the vocational area as well as the academic at advanced levels, assessments must reflect and encourage a far wider range of learning styles than at present. The interests of users can of course differ, between those qualifications for which they expect evidence of a broad understanding of the significance and importance of science in society, and those to which they look for evidence of potential to undertake conceptually demanding studies in future.

It seems clear that the current system of testing and assessment seems to fall far short of meeting accepted criteria of quality in optimum ways. This adds to our concerns about poor reliability a concern about validity, in that reliance on short formal written tests limits very severely the learning aims and styles that can be addressed. In addition, there is clear evidence from a thorough review of the research literature and from our consultations (seminar paper: Harlen; discussion reports: Students' Views) that these regimes de-motivate learners, alienate pupils through the emphasis on grades and competition, and inhibit the development of understanding and of creativity. These effects are bound to inhibit the opportunities for students to engage in those productive ways of learning, notably in respect of meta-cognition, which will equip them as effective learners in the future (seminar paper: Wilson).

Recommendation 8
Any reformed system of assessment should be grounded in a thoroughly researched study of its potential to secure validity, in relation to the need to reflect and encourage a science curriculum that serves the needs and interests of the majority at foundation level, and of the diverse first steps to specialisation at higher levels.

The potential of information technology

The potential of information and communications technology (ICT) to improve assessment practices does need further exploration. The use of 'e-learning' is the issue here, not its use in simply handling and transmission of the data produced by current techniques. Software that can engage a learner in a formative dialogue is an attractive possibility, but given that formative assessment in classrooms is only now being fully exploited and understood, it seems clear that careful development and evaluation will be needed. The prospect is attractive for it can give the learner instant feedback, and can be adaptive in pursuing any interaction in ways that are dependent on the learner's responses. Assessment programmes and packages that are responsive in adapting tasks to the level of the learners' answers might leave teachers more free to deal with individual guidance. They could also help in making better and more flexible use of the time devoted to testing. In addition, the quality of teachers' assessments, both formative and summative, could be enhanced by the provision of banks of tried and tested questions. The new tools that might be developed through ICT could well open up new forms of assessment, forms that are hard to predict. However, partly because they could be so new, suitable resources might best be designed as extra tools to be deployed by a teacher rather than as stand-alone programmes, and trial of their use in the hands of teachers would be essential to establish that they do indeed strengthen and support teachers' learning and assessment practices.

Recommendation 9
The potential of information technology to enable new approaches to both formative and summative assessment should be further explored through development of software programmes which are thoroughly evaluated by trials with teachers.

Valuing and using teachers' summative assessments

The broadening of the styles of learning, ranges of skills, and motivations to learn, that a 14–19 reform must pursue, cannot be achieved unless a wider range of assessment approaches and opportunities is established.

Exams these days are all too alike and yet as teachers we try and deliver the curriculum in the most appropriate way for the types of learners we have. Why does everything have to be tested in the same manner (in Science)? It's only those students who have high literacy skills that do well in science. Teacher

Any broadening of the scope of assessment must involve a new relationship and balance between external formal tests and assessments by teachers, for only assessment by teachers can use evidence shown by students' work over the variety of contexts and time-scales that broader learning opportunities must provide.

Nevertheless, credibility seems to be attached to the current weak and often harmful tools for assessment. Pressures exerted by policy-makers on the development of a broad range of assessment practices to support vocational education have inhibited the development of these practices and done little to raise their status (seminar paper: McKay). The widespread distrust of teachers' assessments is due in part to neglect of the importance of validity in assessment, partly due to the absence of any careful comparison of the potential reliabilities of teacher assessments with those of external tests, and partly due to concerns about copying and plagiarism. The development of credible and manageable methods has not been helped by the pressures that high-stakes tests exert on teachers and their schools. Where teachers' assessments have been given some attention in the system in England, they have been marginalised in relation to the 'league tables' at Key Stages 1, 2, and 3. In GCSE, the extent of their contribution has been strongly limited, indeed cut back. Overall, since the 1970s and 1980s, progress in the area has come to halt, and in some aspects has been reversed.

It is often argued that external testing is needed as a tool to raise attainment standards. The evidence of the effects of testing pressures on raising standards is weak, and there are several studies that show that some of gains in performance on high-stakes tests, which are commonly reported to occur for just the first few years after their introduction, reflect only the development by teachers of the skills of teaching to those tests, rather than any fundamental improvements in learning (Linn, 2000). In addition, because of the inevitable limitations on their length and cost, such tests can in fact damage the very standards of learning attainment that they are designed to improve.

This uncertainty in evidence stands in sharp contrast to the very strong evidence that improvement of teaching methods through strong interactive feedback in the classroom, i.e. through formative assessment, does raise standards of attainment (Black & Wiliam 1998; Black et al. 2004). Although the strengthening of formative assessment is currently given priority in the government's efforts to improve teaching in KS3, there is still room for doubt about the effects of current external test pressures in inhibiting such development. Thus, testing methods should be designed so that they do not restrain such development, and, more ambitiously, should make use of teachers' assessments in ways that can help to raise standards. This need to incorporate teachers' formative practices in a comprehensive review of assessment systems is a significant aspect of current plans for assessment reform in Scotland (seminar paper: Hutchinson)

However, any potential contribution of teachers' assessments to assessment for certification and accountability must be supported by evidence, and procedures, to ensure the quality and comparability of results reported by different teachers and across schools. This requirement has to be met by those methods of checking and inter-calibration, which are summed up by the term 'moderation'. A good example of a method of moderation is one in which groups of science teachers from schools in a local region meet to exchange samples of their students' work to arrive at shared criteria and common standards.

Overall, there is clearly a need for more thorough explorations of both the validity and the reliability of various approaches to designing and interpreting the test data that are commonly used by governments and which command the confidence of a public that does not understand the technical limitations. The research data that shows that current policies are almost certainly far from optimal is rich and varied. All of the evidence that we have surveyed – from research, from practice, and from established theories of measurement and of learning – shows that any reforms should be informed by a thorough consideration of current evidence, about the critical qualities of various approaches to assessment and testing and their close effects on the quality of learning.

Recommendation 10
A new system of assessment should make use both of teachers' assessments and of external measures, and should involve methods of moderating teachers' assessments that can enhance their professional development.

Cautions

However, the published evidence about the use of teachers' evidence for summative purposes gives a very mixed picture. General reviews (Harlen, 2004; seminar paper: Black) and a review focused on teachers in science (Black, 1993) show that several attempts to use teachers' assessments on a large scale have foundered because, either following unfavourable results of evaluation studies or because of public concerns based on anecdotal evidence, confidence in their reliability, fairness and comparability – as between different teachers and different schools – has been undermined. However, there are also examples where both well-researched rigour and

public confidence have been established. Several of the states in Australia rely heavily for the certification of their students at age 18 on teachers' assessments. Some combine them with external tests results, using the latter either as part of the aggregated evidence or as calibration devices at school level (seminar paper: R. Brown). Queensland has relied solely on moderated teacher assessment for over 20 years (Butler, 1995).

Thus, there is evidence that reliability and trust can be achieved, but the research evidence about the pace of innovation and change in the practices of teachers also shows that such achievements take time (see, for example, Black and Wiliam, 2003). Thus, trials of new procedures and methods of professional development, set up on the basis of evidence already available, and carefully developed with school trials lasting at least two years, are essential if robust systems are to be established. Trial and evaluation exercises of this type have to be supported with dedicated funding. One teacher put this point as follows:

There is no "quick fix", the way forward would be a pilot that begins with year 7 and works up through KS3 and KS4. This would give us time to test and refine the system, to make sure that we could ensure that the new system would maintain standards. It should be possible to make sure that over three or four years careful systems could be incorporated into the assessment system that would then be transferable through into Key Stage 4 . . . We are not assessing pupils fairly, we are letting our children down: can we do no better? I hope we can, but we need to make sure that we plan, prepare and test the system thoroughly. Teacher

Recommendation 11
New options and alternative should be tried out and evaluated by practising teachers, and only adopted as policy after such trial and evaluation. This process will require additional funding to both set up and evaluate the innovation. The introduction of new methods should be on a time-scale no shorter than is compatible with the professional development that will be needed. Policy makers should not expect to achieve changes over several months, but should recognise that effective and robust innovations take several years to achieve.

Variety in summative assessment

The argument so far has focused on the need for assessment systems to reflect the priorities of reformed curriculum and pedagogy in science education. It is necessary also to focus on the needs of users, principally the students themselves and those outside schools who use students' assessment results in selection. As they proceed beyond a foundation stage, students ought to have available a variety of routes to qualification, in a system whose structure is reasonably stable and predictable for them. The recent history of the developments in vocational qualifications shows the confusion that a rapid sequence of changes can cause, while also showing that a satisfactory solutions to the problems of vocational education will be hard to find (seminar paper: McKay). They also point to the potential problems of assessment overload that might arise from too heavy a load of coursework assessment. Modular assessment approaches can create similar problems, particularly where requirement for a terminal overall test is added to the requirement for separate assessment of every module. This affects teachers' work, a situation where the curriculum is already too full for them to be able to work through it thoroughly:

We do the modular course because my department believes that our students get higher marks doing it this way. However, so much lesson time is given over to preparing for and taking the modular exams that we end up rushing through the syllabus. Teacher

Students and their parents also feel the pressure:

Because science has modular tests an awful lot of time is spent doing revision in science and my daughter felt that it was going over the same old things time and time again. She came to secondary school excited about doing science but when it came to her GSCEs, she got turned off completely and now she's doing other subjects at A-level simply because they are more interestingly taught. Parent

Nevertheless, there is evidence that many students prefer the modular system (Cerini, Murray & Reiss, 2003).

The history of the development of graded assessment systems in the 1980s is powerfully relevant here. The schemes in science and mathematics, developed in London and elsewhere (seminar paper: M. Brown), enabled students to accumulate credit by progressing through a set of up to 15 levels over the five secondary school years, and the highest level attained at the end of the fifth year earned, automatically, a corresponding GCSE grade. The assessment of attainment at each level used some external instruments, but was mainly based on teacher assessment, checked by external moderation procedures. In the Graded Assessment in Science Project (GASP), teachers assessed students when they felt students were ready to be assessed and were able to use a bank of questions to build tests. Practical skills were assessed by the teacher as were investigative skills, the latter being checked by moderation within networks of teachers. This allowed teachers to reach common understanding on quality and scope in the assessment of students' work. Although the procedures involved teachers in a great deal of extra work, they were popular with teachers, not least because they enhanced the motivation of pupils quite markedly. The introduction

of the National Curriculum was a constraint, but more damaging was Government imposition of a terminal external test which had to count for 50% of the final assessment. This both made the expense unbearable, and undermined the motivation of teachers and pupils as the advantage of no terminal test pressure was removed. Thus the schemes were destroyed because of a belief, not supported by any research data, that accumulated evidence of achievements over five years was not to be trusted on its own, when achievement in a few hours on one occasion in an examination room could be trusted on its own, and had to given equal weight in any acceptable approach.

Overall, if assessment systems for science are to help meet the evident need for a variety of paths aligned to the differing needs and interests, both of those pursuing various avenues of specialism and of those looking for a broader perspective on the subject, then there will have to be a variety of assessments.

First we must recognise the needs of our students. These are diverse and we must recognise that a single comprehensive diet will not do; some students have an interest in basic fundamental science and are mathematically able; others need an awareness of what scientists do, of the tentative nature and limitations of scientific pronouncements, and some grasp of big scientific ideas so that they can function as responsible citizens; still others will be interested in science as a social and cultural activity. Teacher

Given the need for alternatives, we should be concerned with equating of standards between these as much as, and no more than, we are now concerned with equating of standards between (say) physics and biology. It is not possible to both secure the advantage of providing learners with alternative paths to recognition of their achievement, and then require that the paths be closely comparable.

Recommendation 12
A new assessment framework should allow for a variety of routes to qualification – and alternative routes should aim at common standards without a requirement to use identical procedures.

Indeed, the rhetoric of standards can be a source of confusion and can inhibit desirable developments. While the importance of standards and their maintenance cannot be contested, the notion of 'consistent standards' is in fact deeply problematic and the problems are reflected in the contradiction and confusion of media debates (Cresswell, 1996; seminar paper: Baird). A more careful approach is needed, particularly if any significant reforms are to be sensibly managed and fairly evaluated. Thus, for example, an assessment that was a valid reflection of attainment in a narrowly focused course which emphasised the conceptual foundations of topics in (say) physics, could not be valid in relation to a course that emphasised the technological and social implications of the uses of the findings of physicists; the two courses might be judged, by those who valued and understand both sets of aims, to be equally demanding, but the two would require assessments so different that no algorithm could be composed to compare the quality of attainments between the two. It is in any case inevitable that decisions that invest numbers with significance are bound to be based, *au fond*, on the judgements of experts.

Recommendation 13
Comparability of standards between different assessments, designed to be appropriate and valid for the different aims of different courses, must rely on expert judgement and not on uniform and routine procedures which will compromise their validity.

Guidance and motivation of students

One important way in which summative assessments should help students is to guide them in their choices, and to guide the way they go about their next steps in learning. Such purposes are not well served by having a long time interval between the achievement of results which will be rewarded by useable certification. Many students are not well motivated if assessments that 'count' are very infrequent. The value of the graded assessment schemes in producing remarkable improvements in students' motivations is relevant evidence here. Moreover, in a flexible system students should be helped to change a course that they come to find unsuitable for them: they might do this more readily if there were recognised reward for the time that they had invested. A system made of many short components can help learning through feedback of assessment results, if both the curriculum and the assessments are set out in a schedule of progression in the quality, sophistication and scope of learning.

Due to the constraints in the present system teachers find it difficult to help students decide where they are as learners, and the best ways to further their own progress. Teacher

It is not possible for any system strongly dependent on an end-point terminal examination to give continuous guidance to students as they proceed through a course, whereas a system with more frequent medium-term assessments and in which teachers contribute to such assessments can put the teacher in a strong position to advise students about their progress.

As they operate over the years 14–19, assessments should also both reward and reflect progression, and

inform and assist differentiation. We note here that the debate in other countries about issues of measurement, of cognition and of progression seems far ahead of discussions in the UK (seminar papers: Wilson; R. Brown; see also Husen and Postlethwaite, 1996).

Recommendation 14
Summative assessments should be so designed and reported that they serve as a guide to students: this will best be achieved if they reflect a scheme of progression in learning that is embedded in the specification of the curriculum.

Profiles

More generally, the meanings that users attach to, and the inferences that they can make from, assessment results are a central component of validity. When, as is often likely, users are interested in general competence, then assessment results can be aggregated over a range of performances. If a fine-grained profile with separate components of specific pieces knowledge or of skills is required, and if trade-offs across the components are not wanted, it must be recognised that reliability for every component can only be achieved with very great burdens on assessments. Such fine-grained validity should be pursued in appraisal during employment or further study rather than required of this level of the educational sector. Here too there is also a further argument for greater variety in assessment methods, both so that all learners can show their potential, and so that employers can find, in assessment results, evidence for the qualities that they seek.

Recommendation 15
The balance, between aggregation involving compensation between different components, and separate certification of those components in a profile, needs careful design bearing in mind the needs of reliability, and of validity, for all those who use the results.

Manageability, cost and credibility

All the above recommendations must be framed within an emphasis on **manageability**. Our consultations have shown that many in the profession, and many who have researched the effects of current practices, judge that the cost of current high-stakes testing in its effects on classroom teaching, in use of teachers' time and effort and of other resources, is too high and must be reduced. At the same time, as argued above, when the tests themselves are designed to take a short time and to be easy to mark, they can become so invalid that they damage the learning of the subject. Thus there is a strong argument that improvement can only be achieved by placing more reliance on teachers' own assessments, but this does require that teachers have more time for such work.

We'd like to do a range of assessments but there's barely enough time to get them ready for the exams as it is. Teacher

However, there is nevertheless a further and essential criterion, **credibility**, which will need attention to public explanation, given the undue status that the public gives at present to externally set written tests. It is very difficult to evaluate either the manageability or the credibility of any new system in a short time: new procedures are bound to take longer when newly introduced than they will when they are adapted in use and become familiar to users, and the public's opinion about their value will only be established through evidence derived from their implementation rather than from the assurances of hope from their designers.

Recommendation 16
In the design of any new scheme, the need to achieve both manageability and public credibility must be kept in mind, and ample time must be allowed for valid evidence of these qualities to be established.

6 The Way Forward

The aim of this report has been to explore the effects of assessment on the teaching and learning of science. The work of this group would suggest that assessment is entwined inextricably with both curriculum and pedagogy. It is, we have argued, impossible to consider one of these elements, without recognising the consequences for the others. Moreover, the lesson of history would suggest that it is the manner and implementation of assessment that is critical in establishing both the intended and the implemented curriculum. This analysis becomes even more salient in the current context where the outcomes of many assessments are high-stakes for students, teachers, schools and Local Education Authorities. Yet, historically the role of assessment has been undervalued or, even worse, unrecognised. The result is that its form and function have been overwhelmingly of a summative nature and constructed as an afterthought rather than as an integral part of curriculum development. School science has an additional problem in that it deals with a body of consensually well-established knowledge which is too easily presented as a body of facts to be learnt. Assessment in such a context is, likewise, too easily dominated by items that make low-level cognitive demands of recall, application or comprehension rather than evaluation or synthesis.

To give credit where it is due, the national curriculum has attempted to recognise that empirical and investigative work is an essential part of any education in science. In addition, it has acknowledged that the only legitimate and valid means of assessing such knowledge and understanding is through the use of coursework that is assessed by teachers. However, the implementation has led to a situation where whole classes all conduct the same investigation simultaneously. Not unnaturally, the topics for investigation chosen have been those whose nature maximises the potential of the best student achievement: measuring the resistance of a wire, the rates of a chemical reaction, or the rate of osmosis in a potato. This outcome, unintended as it may be, has devalued empirical enquiry in science from a valuable educational experience to a mechanistic and unjustifiable process conducted solely for its outcome where, in so doing, assessment has become decoupled from learning. Moreover, this reductio ad absurdum has alienated many teachers who are rightfully resentful at the requirement to operate and sustain a practice that has little educational value; not least because the commonality of the topics both between and within schools has greatly enhanced the potential for plagiarism.

Thus, for these reasons and others previously outlined, in formulating our view and recommendations about the role of assessment in teaching and learning science we have sought to ensure that assessment works with the intentions of any curriculum and effective pedagogy rather than in conflict. Here our first premise has been that assessment must be a positive experience for students, valued because it validly and reliably reflects their attainments rather than their failings; provides feedback from which the individual can learn; and is yet challenging and thought-provoking.

Recommendation 17
Assessment should be designed to enable students to demonstrate achievement along a sequence of levels of progression in learning, and not to record failures across many different levels.

Currently, despite the original intentions that GCSE should be a unitary examination that would reflect a range of achievements, the focus on the critical C–D borderline has reduced its outcomes to a binary divide between success (grades A–C) and failure (grades D and below). Thus, we welcome and support the recommendations within the Tomlinson Interim report (DFES, 2004) for a foundation-level examination whose singular purpose is to recognise lower levels of achievement, provided of course that it can be shown that schools can provide the teachers time and the other resources needed to add teaching at a foundation level to existing, and other new, commitments.

More fundamentally, the potential to demonstrate achievement will be enhanced through the use of a greater range and diversity of assessment. Improvements could be secured in the present systems by using new and different methods (i.e. different from those described in the supplementary note in the Appendices). A wide range of syllabus topics can be covered in a short time with multiple choice questions, which take less time per topic, and are more economical and reliable to mark, than the structured questions used at present; they can also cover a range of levels of intellectual demand (multiple choice questions are used in some modular examinations at present). A set of these can then free up time for use of a few more open-ended questions to call for more extended and imaginative thinking and writing (the possible gender advantage for boys with multiple choice would be offset by the advantage for girls in open-ended writing). At A-level, there have in the past been excellent examples of the use of timed written tests in the Nuffield Advanced course. These and other examples such as open investigations by students or extended library projects, assessed by their teachers, have all played a significant part in offering varied ways of assessing achievement.

Thus improvements can be secured by drawing on existing tried and tested examples of more valid written and coursework assessments. However, there will still be severe limitations on both reliability and validity, which can only be overcome by the introduction of a wider range of teacher-

based assessment. Our arguments for this are several. First, teachers are responsible for the implementation of the curriculum and its associated pedagogy, two of the three components that are the major influences on any educational practice. Not to permit them any responsibility for the third – assessment – is to ask them to become deliverers whose role is to interpret and enact the curriculum but simultaneously to deny them all influence and choice on a major determinant of their actions. At its worst, it lowers the professional status of teachers by handing to others the agency that is a major influence on their professional lives. In so doing, it also diminishes their sense of self-worth and of engagement in their work. Restoring their role and responsibility must be central to any new initiatives for assessment.

Second, our analysis of the problems of coursework assessment would suggest that is it not the concept itself but the lack of variety that is inherently problematic. When students all do identical investigations simultaneously, extensive opportunities for collaboration and plagiarism are almost an inevitable consequence. If, alternatively, the subcomponents of practical work – the planning and design, the data collection, and the interpretation and evaluation were assessed, there would be two immediate benefits. First, each individual assessment activity would carry less weight, making a smaller and less critical contribution to the overall summative outcome. Secondly, it would permit a wider range of activities for assessment. For instance, we can envisage teachers being asked to select and download one data evaluation exercise from a range of ten. The choice of activities could be regularly changed to avoid the exercise becoming devalued through familiarity. Moreover, there is no reason why assessment should be limited to students' ability with practical work. Students could undertake diagnostic tests items (where potentially their performance could be used formatively); produce a critical review of an article about science; or explain the science in a common piece of technology. One teacher took this idea further:

Some GCSE awards might be assessed by methods which look similar to those we have now (though with more generally descriptive criteria); others might be assessed through methods more akin to those used in English or art examinations or by the sort of oral examinations used by modern language teachers, standardised through visits from peers. Such systems would be expensive: the pay-off would come through more motivated learners, more appropriately educated learners and more creative teachers! Teacher

Such diversity would measure a wider range of student competencies, making the assessment more valid than at present. Students could be expected to undertake several such assessments during the period of their course from which they construct a portfolio of work as evidence of their learning and achievement.

Recommendation 18
The validity of assessment in science would be improved by the use of a greater range of styles and modes of assessment. This applies both to formal examination and coursework.

Teachers of science must be able to offer an education that enthuses them and, as a corollary, their students. Given that assessment is the critical determinant of both curriculum and pedagogy, it is, therefore, vital to restore some sense of ownership of the means by which students are assessed. Only this will restore to teachers a measure of responsibility that is the legitimate expectation of a trusted professional. Variety, of itself, will not address this issue. Rather, we believe teachers should be offered the opportunity to engage in the process of developing new approaches to assessment themselves.

We envisage a process where groups of teachers working in collaboration produce items and approaches to assessment that are tested, trialled and moderated by the group. Although this would take time, and would require a measure of financial support, what would evolve is a community of practice with a greater degree of professional competency and expertise, among teachers of science, of valid and reliable modes of assessment. This knowledge is essential to develop a better and more appropriate use of assessment both formatively and summatively. Some level of competence already exists among those teachers of science working in vocational education. The model of assessment piloted by the new 21st Century Science at GCSE incorporates attempts to break away from the malaise whereby rich learning practices are inhibited by a curriculum narrow in range, crowded in content, and reinforced by pressures of narrow assessments. It aims to do so by offering a greater range of the assessment tools and trusting teachers to be able to make assessments alongside their teaching (seminar paper: Nicholson).

In addition, there are many lessons to be learnt from previous approaches that made extensive use of teacher assessment, such as the graded assessment projects from the 1980s (seminar papers: M. Brown). Although it is accepted that education is so important that its outcomes must be measurable, measurement alone cannot produce high-quality outcomes: the quality of teachers and their professional commitment to what they are asked to offer pupils as part of their daily fare is the *sine qua non*.

At Key Stage 4 they (students) lose a lot of enthusiasm; that could partially be as a result of the content of the syllabus but also because of the way they are assessed through the course. Unless teachers value and truly believe in what they are doing in classrooms and think that their actions are going to have a long-lasting impact and importance to their students, there is little hope of inspiring students into furthering their studies in science. Teacher

Sustaining teachers' enthusiasm for the science they teach can only be achieved if they have some stake in determining what is sufficiently significant and important to be measurable. But outcomes of high quality in science education will be dependent on a new approach, one designed to achieve optimum linkage between a *curriculum* reformed to emphasise new aims; *teaching and learning* practices that can achieve these aims; and *assessments* that reflect and support these aims and practices. The complexity of the interactions between these elements is such that synergy can only be achieved by close involvement of practising teachers in new developments. This argument leads us to our final recommendations.

Recommendation 19
New courses in science education should reflect aims that are relevant and appropriate to the students, should support quality approaches to learning, and should be mirrored and supported by reformed systems of assessment. Science teachers should play a leading part in developing and integrating these strands of reform, and in the assessment procedures that should emerge.

In any reformed system, it needs to be recognised that any assessment system cannot be perfect in its original design, and that changes in science, in students and in the needs of society, will call for evolution. The design of assessment and curricula should foster ways of encouraging and accommodating such evolution, and of supporting creative teachers, perhaps in partnership with other groups, in piloting new ventures. One key aspect of such support is that the assessment system must allow for new pieces of schoolwork to 'count' for assessment purposes.

Recommendation 20
A new system should be so designed that innovation and renewal can be a continuing part of its operation: groups of teachers, with others, should be encouraged and supported in developing new innovations, and the system should allow for a small proportion of the public assessment measures to be used to assign credit for assessment of the innovative work.

Appendices

The following accounts are of two kinds. Some are brief reports of the various papers presented to the project's seminars by invited experts. Others are summaries of the methodology and the outcomes of discussions that took place either in the three seminars or in the open meetings. In those cases where a fuller paper, or other account, is available, this fuller version will be available on the King's College London web-site.

Short reviews of the papers presented at the three Seminars

1. Would the *real* gold standard please step forward?
   Paper presented at the first seminar by Jo-Anne Baird Assessment and Qualifications Alliance

2. Tests and assessments: purposes and quality.
   Paper presented at the first seminar by Paul Black, King's College London

3. The use of teachers' assessments for public summative certification.
   Paper presented at the third seminar by Paul Black, King's College London

4. GAIM and GASP
   Paper presented at the third seminar by Margaret Brown, King's College London

5. Assessment in science education in Australia.
   Paper presented at the second seminar by Roger Brown

6. Teachers' views of assessment and learning in the post-14 age group.
   A summary of papers presented by Simon Carson, Suzanne Parkyn and Roy Wagner

7. Testing, motivation and learning.
   Paper presented at the first seminar by Wynne Harlen, Assessment Reform Group

8. Assessment, testing and reporting 3–14: A Scottish perspective.
   Paper presented at the second seminar by Carolyn Hutchinson, Scottish Executive

9. The recent history of vocational qualifications in schools and colleges.
   Paper presented at the second seminar by David McKay, QCA

10. Assessment in 21st Century Science.
    Paper presented at the second seminar by Peter Nicholson, York University

11. The BEAR project.
    Paper presented at the second seminar by Mark Wilson, University of California, Berkeley

Reports

12. Report on the Methodology

13. A summary of the Open Meetings

14. A summary of the parents' focus meeting

Supplementary paper

15. The nature of examinations at GCSE and A-level

Paper 1

Would the *Real* Gold Standard Please Step Forward?

Paper presented at the first seminar by Jo-Anne Baird, Assessment and Qualifications Alliance

Debate about public examination standards has been a consistent feature of educational assessment in Britain over the past few decades. The most frequently voiced concern has been that public examination standards have fallen over the years: for example, the so-called A-level gold standard may be slipping. The paper considers some of the claims that have been made about falling standards and argues that they may reveal a variety of assumptions about the nature of examination standards and what it means to maintain them.

The notion of consistent standards may be founded on criterion-referencing, or on norm- (or, more strictly, cohort-) referencing, or on requiring that standards be the same if *all* factors that are predictive success for the candidates concerned are identical, or on a sociological basis that a due process for determining standards (e.g. by experts' judgments) has been followed. Each of these perspectives is examined in detail and it is argued that, because people disagree about these fundamental matters, examination standards can never be maintained to everyone's satisfaction.

The practical implications of the various co-existing definitions of examination standards and their implication for the perceived fairness of the examinations is considered, but it is concluded that the adoption of a single definition of examination standards would not be desirable in practice. It follows that examination boards can legitimately be required to defend their maintenance of standards against challenges from a range of possibly conflicting perspectives. This makes it essential for these boards to be open about the problematic nature of examination standards and of the processes by which they are determined.

*This paper was based on a paper of the same title by Baird, Cresswell and Newton in Research Papers in Education, **15**(2), pp.213–229, 2000.*

Paper 2

Tests and Assessments: Purposes and Quality

Paper presented at the first seminar by Paul Black, King's College London

Designing the optimum system

A perfect solution to the problems of assessment and testing is not possible. One can only try to achieve the optimum resolution of the various conflicts, between the different purposes, different criteria of quality, and the constraints on time and resources.

Purposes of assessment and testing

The three main purposes of assessment, for learning, for certification and for accountability may require different types of evidence of students' achievements, and/or different ways of analysing such evidence. Assessment for learning is frequent, informal, and an integral part of teaching and learning, whereas assessment for certification has formal constraints because the results must have public credibility. The purposes of accountability might best be served by sample surveys using a range of tests with different samples, but this purpose also calls for a collection of background data to assist in interpretation. There may be useful synergy between the information collected for feedback in promoting learning and use of the same information to serve the other two purposes, but there may also be conflict insofar as the instruments used for certification and accountability are not designed to yield information about learning needs.

Criteria of quality for tests and assessments

There should be concern about the poor *reliability* of short tests set on formal occasions, yet there is little dependable evidence about the chances of candidates being wrongly graded by such tests. Such evidence as exists indicates that as many as 30% of the entry may be wrongly classified, mainly because of the inherent variability in students' performances, between different questions and from one occasion to another. Poor *validity* is also serious concern. Some valued learning outcomes, particularly those now being developed to broaden the aims of science education, may not be reflected in test results because they cannot be assessed within present constraints.

Teaching learning and assessment

Better assessment for learning by teachers could raise standards, and weaknesses in both reliability and validity could be reduced by better use of teachers' own assessments. However, while there is evidence that such improvements have been achieved, it is also clear that extensive professional development work with teachers will be needed to ensure that the rigour and comparability of teachers' assessments can command public confidence.

Paper 3

The Use of Teachers' Assessments for Public Summative Certification

Paper presented at the third seminar by Paul Black, King's College London

This paper explains the difficulties and constraints there are in using assessment ideas in a general manner and the need to decide on purpose to hone tools effectively. The formative assessment expertise of teachers on the King's–Medway–Oxfordshire Formative Assessment Project (KMOFAP) was greatly enhanced, and these teachers were well able to describe the attributes of their learners. However, no work was done to translate this knowledge of their learners into defensible numerical scores or grades. In fact, these teachers used similar continuous summative testing as other teachers, mainly because time pressures prevented them devising assessments of their own.

There were cautionary tales of various studies that had tried to incorporate rubric-driven instruction as a means of improving reliability and allowing comparability of different teachers assessing. It was felt that teachers need support in this area such as the provision of test instruments that they can select from and use at their discretion.

What was suggested were possible ways in which effective teacher-led assessment could be developed:

a) portfolios in which perhaps learners have the opportunity to select work;
b) collections of data from some pieces of the normal classroom work to include tests and homework;
c) data derived from tests drawn from external question banks;
d) graded assessment schemes which combine some of the above in a system.

The decisions and problems that have to be faced were also listed:

a) teacher workload;
b) comparability of validity of different teacher assessments;
c) guarantees to prevent cheating/plagiarism;
d) whether teacher-led assessment assesses the same as external examinations or each assesses particular attributes of learning;
e) manageability and cost.

Paper 4

On GAIM and GASP

Paper presented at the third seminar by Margaret Brown, King's College London

The Graded Assessment Projects were a suite of schemes developed during 1983–90 with the aim of providing formative assessment throughout the 11–16 age range which was convertible into GCSE grades at age 16. The partners were the Inner London Education Authority, the London-based Examination Boards and King's College London, but considerable support came from other LEAs throughout the country and from the Nuffield Foundation. The subjects covered were Science (GASP), Mathematics (GAIM), Design and Technology (GACDT), modern and community languages (GAML) and English (GAPE).

One common feature across all subjects was a level structure that would, in most cases, allow pupils to attain, at the rate of one level per year whatever their starting point, an assessment that was later incorporated in a modified form into the national curriculum. Levels were defined by criteria, which were characterized as far as possible as conceptually based skills, relating to both content and process. Being aware that separate criteria might fragment assessment and, more importantly, teaching; in all subjects holistic open tasks were, therefore, a key part of the assessment methods. Tasks and their associated levelled assessments were developed and trialled in the pilot schools, where they were mostly completed in lesson time, avoiding the problems of reliability with GCSE coursework.

Assessment of criteria was more flexible; they could be assessed as part of the open tasks, by more closed tasks and tests, by peer assessment, or by classroom observation and questioning. In some cases, cross curricular methods were used, usually with pupils reporting that they already satisfied criteria in other subjects and producing relevant evidence. Schools starting on the schemes often found it easiest to start with more routine test questions selected from a bank provided but became more sophisticated as they became more familiar with the criteria. In some subjects a gap was required between teaching and testing so that only that knowledge which was more permanent was assessed.

All graded assessment schemes aimed for a high level of student participation, with students, who had ownership of their record sheets, being encouraged to provide evidence of satisfaction of any criteria to their teachers, and tackling creatively challenging, open activities, knowing the broad criteria for success.

Certificates, either as cross-level profiles showing areas of strength and weakness, or as level certificates, were available to schools to issue when ready, and GCSE grades were awarded at the level reached at age 16 (or when required, but normally only on one occasion). Lead teachers attended training sessions nationally and in local cluster groups, and local assessors appointed by the examination board visited schools regularly to provide both moderation of standards and support.

Teachers generally reported that they enjoyed operating the schemes and profited from the professional development, but the motivating effect of the scheme on pupils was the overwhelming benefit of the schemes. Initial problems of over-complexity and workload were solved during piloting. Over 1000 schools were using GASP by 1993; numbers on GAIM were limited as it had to be a pilot GCSE as a waiver of the final examination requirement had to be granted; nevertheless there were over 150 schools involved. Costs were calculated so that they were spread through the age range but that the total was the same as that for a GCSE. High inter-subject correlations suggest that the results were reliable and the examination board was satisfied that the awards were appropriate and fair. However, soon after the GCSEs started to function, there was a national announcement in 1991 that final externally marked and set examinations would account for at least 80% of GCSE marks. This meant that the conversion of formative to summative GCSE grades was no longer possible and the numbers gradually fell off, although some teachers are still using the materials. Thus a system that appeared to satisfy the requirements of Tomlinson, and to be rigorous, to encourage a more open and creative curriculum, and to encourage engagement among students, was abandoned without having been formally evaluated. Maybe it is time to resurrect graded assessment, with the modifications, e.g. more use of computers, which would be needed to update it.

Paper 5

Assessment in Science Education in Australia

Paper presented at the second seminar by Roger Brown

The system in Australia differs significantly from that in England and Wales in that Australia is a federation of relatively autonomous states. There does exist a federal curriculum framework, which sets the goals of schooling but no national assessment system.

Assessment takes place at age 16 but only in New South Wales is there an external examination. Other states use a system of internal assessment by teachers where it is left to the professional judgement of teachers to devise an appropriate assessment system. Eighty per cent of students will stay on at school until age 18 and approximately 65% go on to some form of higher education.

Hence each state-based examination board offers its own form of assessment, which differ from one another. Three examples are given below of the kind of assessment used in science at age 18/19, the age at which the overwhelming majority leave school.

Victoria	Each semester, in the final year of schooling (age 18), there is an examination in each of the science subjects consisting of a combination of multiple choice and short answer questions. The external element contributes 66% of the mark and rest is based on school-based assessment.
New South Wales	The school-based assessment and the external examination at the end of each semester have equal weighting. Scores are reported in terms of a moderated school assessment mark, a Higher School Certificate (HSC) Mark, which is an average of the external and school-based mark, and their performance band which shows where their HSC mark lies in relation to all other candidates.
Queensland	No external examinations are held and all assessment is conducted internally and then cross-moderated in schools. This system has now been in operation for over 15 years. Essentially, students are ranked internally by using a system of subject achievement indicators to determine their relative achievement within the school. Subject achievement indicators for each school are then ranked by a system of moderation.

Although there is some variation in the form of school-based assessment, the general, most states use a mixture of practical-based reports and investigations, which are moderated externally before finalising the grade.

In Victoria, political concern about school-based assessment has led to the use of a General Achieved Test (GAT) as a means of ensuring consistency between schools. The results of this test are not reported publicly. Instead, they are used to compare the marks attained on this test with the marks and distribution of school-based assessments. Where there is a significant mismatch between schools and the predictions of the GAT test, the assessments of the schools are moderated and investigated in some detail. This mechanism is a means of maintaining public confidence in school-based assessment.

Likewise, Queensland has a Queensland Core Skills Test which tests the 49 elements that are common to the curriculum. The test is again used as a means of checking the reliability of schools' assessment of students' work.

Paper 6

Teachers' views of Assessment and Learning in the Post-14 Age Group

A summary of papers presented by Simon Carson, Suzanne Parkyn and Roy Wagner

The options open in choosing examinations at 16 have narrowed and teachers feel driven to choose the one for which they think they can help their students get the highest grades. One teacher criticized the whole system because it is based on a specification that requires students to memorise and repeat facts about scientific knowledge that are of little interest or relevance for them. Many questions require little more than substitution of numbers in a formula, and the atomistic approach of many short questions masks the process of scientific discovery and reduces science to a fixed body of facts. The drive for teachers to maximise scores is thereby in conflict with their wish to help learners explore the intellectual demands of science.

However, the most severe criticism was reserved for the assessment of scientific investigations. The rules to which teachers' assessments must conform, and the further constraints imposed by external moderators, have reduced the work to a process of getting students to jump through clearly defined hoops. There is little variety in the tasks set: one in widespread use is an investigation of how the length of a piece of wire affects its resistance: this is popular because the results are reliable, repeatable, and always provide a simple linear relationship. The results are no surprise to students – and they have no interest in them. The work has become a travesty of scientific enquiry. The stereotyping that assessment has imposed on these investigations has also led to widespread opportunity for plagiarism and cheating.

The positive plea is for development work to be undertaken with the aim of handing over more, or even all, responsibility assessment of their students' achievement to teachers – who know far more about their students than can ever by 'measured' by any feasible external test system. It is recognised that long-term professional development, in which the public are involved and kept informed, is necessary if this is to be acceptable. There was also plea for greater variety in the methods and styles of assessments and testing so that they can reflect the variety of strengths and interests of students.

Paper 7

Testing, Motivation and Learning

Paper presented at the first seminar by Wynne Harlen, Assessment Reform Group

Do tests have a positive or a negative effect on pupils' motivation?

Some believe that tests motivate pupils to work harder and more effectively, while others argue that they give rise to stress which is de-motivating. The paper reviewed the research evidence bearing on the links between summative assessments and pupils' motivation to learn, and discussed the implications of this research evidence.

Research findings about the impact of tests

The overall impact of formal testing is a negative one. For example, the introduction of national curriculum tests lowered the self-esteem of those who did not perform well and their motivation to learn was also reduced – and the gap between low and high achievers was enhanced. Test performance was more highly valued than what was being learned. Among older pupils, the low achievers showed more anxiety, resentment, cynicism and mistrust of external tests, with girls showing more test anxiety than boys. Under the pressures of 'high-stakes' testing, teachers become expert at teaching to the test, to the detriment of teaching for understanding or for developing creativity.

Implications for the work of teachers

Teachers can reduce the negative effects of tests by helping pupils to understand the processes and contexts of the construction and marking of tests. They can also help them both by leading them to take more responsibility for, and control over, their learning, and to develop a clearer understanding of the aims of their learning, so that overall they become more confident and effective as learners.

Implications for policies at school, local and national levels

Professional development work is clearly needed. Those responsible for local and national assessment policies should re-think their policies in the light of the evidence of the negative effects of current tests, which arise both from their limited validity and the 'high-stakes' pressures that they impose on schools to the detriment of good learning and of the self-esteem of many pupils.

This paper was based on a booklet with the same title published by the Assessment Reform Group

Paper 8

Assessment, Testing and Reporting 3–14: A Scottish Perspective.

Paper presented at the second seminar by Carolyn Hutchinson, Scottish Executive

This presentation was based on the work of the Scottish development programme which has been running ten national assessment projects over the past two years. One of the most successful of these has been 'Assessment is for Learning', which has been running several pilot projects in schools in Scotland since 2002. The outcome of this work has been independently evaluated and has led to a set of proposals for revising the current system to improve the contribution that assessment can make to learning.

Key Issues emerging from the Assessment programme are that assessment has four common uses:

- monitoring national standards and the quality of educational provision;
- a summative role to select and certify student attainment and to provide guidance and feedback to students on their attainment;
- a diagnostic role to identify students' learning difficulties;
- a formative role to identify what needs to be done to improve student learning.

In the present system in Scotland, the emphasis tends to be on the first of these two purposes, with teachers paying undue attention only to those components that can be measured by national and standardised tests. The diagnostic and formative roles of assessment are undervalued and underused. Parents play little active part in assessment. Consequently, the report makes three recommendations to change the system, which are currently under consultation.

1. Replacing reports with Annual Progress plans. These differ significantly from the more familiar annual reports in that, as well as summarising progress, they would be supplemented by a personal learning plan that identifies the steps that need to be taken to make further progress. As such, the information in the document will enable parents to engage more in their children's learning.

2. Simplifying the system of assessment and ending the current system of national tests for 5–14 year olds. Although the Scottish system was designed as one that would enable teachers to test when ready to provide supplementary information and confirm teachers' judgement, in the ten years since its inception it has become subverted. Teachers feel that they must test all pupils at the same time and use these tests as a means of ranking pupils. Another consequence is that teachers find themselves teaching to the test but, inevitably, some pupils are not be ready. The consequence is that, compared with their peers, they are likely to fail and the outcome will be demotivating. Such tests focus on a limited subset of pupil competencies and often summarise pupil attainment in a single number or letter. Such information does little to inform parents about how well their child is learning and has virtually no formative or diagnostic value.

The Scottish proposal is to introduce a set of items that will be randomly selected from a bank online. Teachers will be unable to select items from a catalogue or peruse them in advance, giving less incentive for practising.

Changing the method of selecting the tests will not, of itself, generate any change in the underlying system. Rather it is proposed to move to a system of sample-based surveys to monitor national performance akin to the TIMMS study. This would permit a greater range of assessments covering a greater range of each subject and in more detail. Such a system would also permit the inclusion of practical tests. Surveys would be administered 'unseen' to pupils without preparation and without practice, thus eliminating 'teaching to the test'.

Paper 9

The Recent History of Vocational Qualifications in Schools and Colleges

Paper presented at the second seminar by David McKay, QCA

The history of vocational qualifications in science in England is not a simple picture. Following the introduction of GCSEs in 1986, there has been an ongoing attempt to establish a satisfactory system of vocational qualifications and applied courses in science. Post-16, the major development was the introduction in 1993/94 of the twelve unit Advanced General National Vocational Qualification (GNVQ) and a six unit (one year) Foundation and Intermediate Qualifications. The Advanced GNVQ was intended to provide a vocational alternative to GCE A-level with the option of moving on to higher education or employment.

Foundation and Intermediate courses were designed as one-year, full-time programmes, to be more motivating than repeating GCSEs, each award providing the option of moving to a higher level of qualification, to new areas of study, or into employment and training. These then offered three levels of qualification, each attainable as a pass, a credit or a distinction.

The original concept of assessment for these examinations was one based on a portfolio of work aimed at demonstrating 'mastery' against set criteria and internally assessed. However, the Department of Education insisted on an element of external testing and grading, generally using multiple-choice items. These two elements have existed in tension and still do to this day. Each test and assessment has to be passed to gain the qualification, and the essential aim was to create a distinctive approach to learning encouraging more young people to achieve, and to recognise that achievement

through an appropriate assessment system. The use of a rigorous but differentiated system with a wide range of levels was, therefore, an attempt to gain user confidence. In addition, the system also required the assessment of core (later key) skills and grading these against generic criteria of 'planning', 'information seeking and handling' and 'evaluation'. A perhaps inevitable outcome of such a system, though, was that the resulting assessment system was considerable.

Initially, these qualifications were warmly received by further education colleges but growth slowed as a consequence of a series of problems. In the main these were:

- the maintenance by awarding bodies of their 'own brand' of qualifications;
- reservations about the vocational relevance of some content;
- concern about the substance and standard, and hence the worth of the vocational qualifications in comparison to GC(S)Es;
- concern about the trivial nature of some of the 'mastery' tests which were mostly simple multiple-choice items testing recall;
- doubt about the rigour and consistency of assessment and grading decisions;
- problems arising from the unmanageability of assessment arrangements.

The result was the establishment of a review undertaken by John Capey (Principal of Exeter College and Chair of NCVQ's GNVQ committee), which recommended various steps to improve manageability and rigour. In the event, the Dearing report overtook the Capey Recommendations and the Curriculum 2000 reforms saw the new Vocational A-levels replace GNVQs. However, the focus of Curriculum 2000 was A-Level, and retaining the facility to teach the constituent units of the curriculum in any order has meant that each unit has to be pitched at A2 standard and that the VCE has no equivalent to the lower standard of the Advanced Supplementary (AS) level. Consequently, the Qualifications and Curriculum Authority is now working to change the model yet again to parallel the AS/A2 structure for the GCE for launch in September 2005.

Paper 10

Assessment in 21st Century Science

Paper presented at the second seminar by Peter Nicholson, York University

21st Century Science is the new science curriculum currently (2003–5) under development by the University of York Science Curriculum Centre. This form of science course is a product of the report Beyond 2000: Science Education for the Future. It is an attempt to resolve the inherent tension of a science curriculum which both needs to educate the future citizen (and non-scientist) about science, its cultural achievements and its role in society; and also educate the future scientist. This is done by offering a GCSE science course in two parts. First a general Core Science course in all three sciences, which may be loosely considered as a course in science for citizenship or public understanding. This course is equivalent to one GCSE and taken by all students. Second, additional optional courses in either Additional Core Science which is very similar to traditional science courses or Additional Applied Science. The latter offers a pre-vocational study of how science underpins a range of areas of professional activity and develops general competencies useful for the world of work.

Examinations for each of the three courses consist of a mix of written examinations and coursework assessment. The percentage weighting for coursework assessment will be higher than current science GCSEs. Of interest to this report and the issues addressed by the seminar was the nature of the coursework assessment, which will be significantly different to traditional forms of coursework assessment.

a) In the Core Science, students will be required to carry out two pieces of coursework – a case-study and data analysis exercise. The case-study will require students to identify a science-related issue in the media, to explain some of the background science, to consider the risks and benefits associated, and to come to a considered view about the issue. These reports will be internally marked and then externally moderated. The aim of the coursework is to develop an understanding of how science is reported in the media and to develop their ability to review critically the validity and reliability of such reports.

   The data analysis exercise will require students to collect primary data by using a practical procedure. The focus of the assessment, however, will be the students' ability to analyse and evaluate the data and the limitations of the empirical techniques.

b) In Additional General Science, students will have to undertake two additional pieces of work: an investigation and an open-book assessment. The investigation is akin to the kind currently required for all science GCSEs and designed to develop students' ability to identify a clear and manageable question for inquiry, how to choose equipment, to use it appropriately, to make suitable observations and measurements, and to evaluate and interpret the data.

   For the open book task, each student will receive stimulus material consisting of several extracts from scientific papers, magazines, newspapers or other sources that will contain data and information on a specific science topic. In addition, each student will receive a question and answer book and will then have two weeks to study the topic using any additional sources of information and write their answers. A substantial part of lesson time will be devoted to this task over the two weeks. Work will be internally assessed and externally moderated.

c) For the Additional Applied Science, students will be required to undertake an investigation. This will require first-hand experience of the problems of collecting valid and reliable data and provide context for developing student problem-solving skills. The work will be assessed under five headings: the ability of the student to a) devise a strategy to investigate the problem; b) collect data; c) interpret data; d) to evaluate and draw conclusions; and e) to present the findings. Again, coursework will be internally assessed and externally moderated.

One of the issues of concern is the potential for collusion and plagiarising work. However, one proposed mechanism for addressing this is a requirement that not all students undertake a common assessment, minimising the possibility of collaboration. One response was that collaborative work was an essential element of science and that we should consider developing mechanisms by which such work could be undertaken.

Paper 11

The BEAR Project

Paper presented at the second seminar by Mark Wilson, University of California, Berkeley

This paper reported on the approach to assessing, interpreting and monitoring student performance developed by the Berkeley Evaluation and Assessment Research (BEAR) Center. Its aims are to provide a set of tools to:

- reliably assess student performance on central concepts and skills;
- set standards of student performance;
- provide feedback on student progress for a range of audiences.

Fundamentally, this project sees the function of classroom-based assessment as a means to generate high quality evidence produced from a system, which embeds assessment as a central feature of learning. Their work is based on four principles:

(a) A developmental perspective which seeks to measure the progress of the student. Such an approach requires repeated measurement of student progress using a range of different assessment methods.

(b) Classroom-based assessments that must generate evidence of quality. This means developing items that match the standards of reliability and validity of standardized tests and not using flawed questions.

(c) Matching what is taught with what is assessed. Assessments must be designed so that teachers develop the competencies that are the real goals of educational reform. As it is, too many tests overemphasise the recall of memorised information and not the higher-order skills of evaluation and synthesis.

(d) Teachers must manage and implement the system of assessment. National or state-wide systems of assessment cannot provide the need for immediate feedback to manage instruction and monitor. For this to happen, teachers must be:
  - involved in the process of collecting and selecting student work;
  - able to score and use the results immediately, not wait months for the scores to be returned;
  - able to the implications of results for future instructions;
  - able to take a creative role in designing and implementing assessment.

The paper illustrated the development of these principles and ideas through a set of progress variables, called *Perspective of Chemists*, which embody the progression in understanding from novice to expert. In addition, these incorporate the National and California State Science Education Standards. Such a calibrated scale map of the growth of students enabled teachers to track student progress and to cross easily between standards and assessment.

To gain quality evidence, this group have used an approach called item response modelling, which can locate a student or a class along a progress variable. Such techniques generate reliability coefficients, enabling inter-rater comparisons. Such a system was also used to generate progress maps, providing feedback to both students and their parents.

Assessment tasks need to reflect the range and styles of instructional practices in the curriculum. Hence, the instructional materials were developed simultaneously with assessment tasks. Doing so enabled the creativity of curriculum development to be embedded also in its assessment and forced the discipline and struggle of generating valid assessment into the design of the curriculum.

To enable the system to be managed by teachers, scoring guides accompanied by concrete examples were developed which defined the performance criteria. These were accompanied by assessment blueprints, which assisted teachers to decide when assessment was appropriate.

This system was implemented in a science course developed by the Lawrence Hall of Science and tested with a large sample of students and a comparison group. Students following the Bear system made significantly greater learning gains ($p < 0.05$) than the comparison group. The results provide evidence that considerable learning gains can be achieved by (a) closer attention to assessment concerns at the classroom level, and (b) a more systematic approach to the gathering and interpretation of assessment information.

Paper 12

Report on the Methodology

The rationale for a focused seminar series as our approach to investigate the issues relating to the 14–19 assessment of science learning was that it enabled a detailed discussion and comprehensive investigation of the many factors, which other forms of approach, such as questionnaires, would be unable to elucidate. This approach proved successful with the Beyond 2000 seminar series in producing a report that informed and influenced government, schools, researchers and the public about the types of curricula that would benefit youngsters in responding to, and making informed decisions on, issues raised by science in our modern world. We wished to make an informed and useful contribution to the current discussions relating to the Tomlinson reform of the 14–19 phase, and possibly help foster change in the ways in which science is currently assessed and taught. To achieve this end, we recognised that we would need to delve deeply into the issues so that ways forward can be sought that are both effective and pragmatic, and hence, to arrive at recommendations that will be acceptable to the many parties who express an interest in this area.

The approach that we took in the three seminars was to commission lead articles from experts in the areas of science, science education and educational assessment and also contributions from the three practising teachers on the seminar group. These papers were read by the seminar participants before each seminar and then presented at each seminar by the author. Each paper was then given a prepared response by one of the seminar group before the group split into two smaller groups for discussion of issues arising from each paper. This was then followed by a feedback session from each of the subgroups before a final discussion of the paper by the whole group. The presentation and response were recorded in note form by two researchers. Each of the subgroup discussions was also documented by a researcher, with the feedback and final discussion again noted by two researchers before we passed onto the next paper, where the process was repeated.

At the end of each day, the group spent some time reflecting on the group of papers that they had discussed, and issues that arose were duly noted by one of the researchers. Some participants sent in further reflections after the seminar.

Each of these data sets were accumulated and re-read by one researcher, who had been present at the seminar. Ideas, issues and responses were categorised and coded. The codes were matched, challenged and interrogated from paper to paper in a search for main ideas, contradictory evidence or thoughts and variation for different interest groups. These were then discussed at regular meetings of the King's College team both to validate the data and feed into our emerging ideas and to plan for the following seminar and open meetings.

The open meetings were used to discuss and test out emerging ideas and also to collect further data. One researcher took main points of whole group discussions. Because of the wide range of participants from school students and teachers to university entrance tutors and examination board officers, it was decided to organise some sessions where groups from similar backgrounds (e.g. school students) could meet, and at other times to use mixed groups. In this way, we hoped that everyone would have the opportunity to have their say but also enable different stakeholders to hear and respond to the ideas from others. These discussions were self-reported on summary sheets. Again the notes were categorised and coded and cross-referenced with the data from the two seminars.

As well as the two open meetings in London and Sheffield, we held open meetings at the British Association of Science Education Conference in October 2003 and at the Association of Science Annual Conference in January 2004. Although we did not take detailed notes in these two sessions, a report on each session was written shortly after each and treated in the same way as the notes from the seminars and open meetings.

Paper 13

A Summary of the Open Meetings

All representatives at the Open Meetings felt that current science lessons in schools were insufficiently stimulating, failing to encourage an interest in science. Students at the Sheffield meeting voiced strongly that, although they would like interesting science lessons, they saw their reason for being at school was to obtain good examination grades and many felt that 'teaching to the test' was a good approach to take to learning science, contrary to the views of most of the other groups represented at the meeting.

Teachers at the open meetings questioned the validity of the current assessment schemes. They perceived the main difficulty as one of interpreting annual results because comparability is not straightforward when each year different papers are set, cohort sizes may alter and curricula change. Examination boards and Government agencies have tried to deal with this by reducing flexibility in the types of assessment available. Although this may improve comparability, there was concern that this led to under-examination of important scientific skills, such as problem solving, scientific enquiry and communication of ideas. Some teachers also stated that it was inappropriate to test solely by written examinations as this disadvantaged some students' ability to demonstrate their scientific understanding if they had low literacy skills.

One solution for more comprehensive testing of science skills suggested was for teachers to take on a larger role in assessment. Teachers are best placed to produce valid assessments because they also control how science is going to be learned in the classroom. Although there may be some training needs in helping teachers find suitable activities to use for assessment, it is clear that validity would be high if the major part of assessment was performed as ongoing work alongside the learning rather than attempting to test those parts of the learning that could be assessed by written examination papers. That such an approach would clearly have consequences for teacher workload was strongly voiced by teachers at both open meetings.

However, the most severe criticism of the current assessment schemes was reserved for the assessment of scientific investigations at KS4. The rules to which teachers' assessments must conform, and the further constraints imposed by external moderators, have reduced the work to a process of getting students to jump through clearly defined hoops. This has led to little variety in the tasks set and the reliance on a set of tried and tested practical tasks in which the results are no surprise to students. That such work has become a travesty of scientific enquiry was agreed by all groups at both open meetings. The ritualistic approach that such assessment has imposed on these investigations has also led to widespread opportunity for plagiarism and cheating. This caused concern for some teachers and headteachers, and brought into focus how it would be possible for the exam board personnel questions to monitor and overcome problems of this nature.

Paper 14

A Summary of the Parents' Focus Group Meeting

The parents' focus group consisted of six parents who had at least one child in the 15–17 age group, and so their children were either working towards GCSE that year, or had taken GCSE in the past two years. Three of the parents had children who had opted for one or more science subjects for GCE A-level.

Discussions began with statements from several parents that they believed that external examinations convey objectivity and reliability and as such are preferable to internal assessment tasks devised and administered by teachers. However, as the discussion progressed, all but one parent stated that their children had found the frequency and approach to assessment in science at GCSE stressful. All children had taken modular science courses and whereas the parents could see some advantage in assessing during the course, they reported that each module test was like preparing for a full examination, leading them to conclude that science seemed to over-assess in relation to other subjects.

There was concern from two of the parents about whether teachers and schools could be trusted to fairly assess children when so much was at stake both for the school's reputation, financial rewards and status within the local community. It was generally thought that teachers already have high workloads and that more teacher-led assessment may not be received well by teachers.

Two of the parents spoke about how their children had been excited about science and saw it as a possible career when first at secondary school. However, the focus on learning science facts and a diet of undertaking and writing up practicals in a formal way had resulted in their children deciding on other subjects for GCE A-level study, even though they had done well in their science GCSEs. It was felt by these parents that science at the upper end of secondary school was preparation for written examinations to a much greater extent than other subject areas.

Paper 15

The Nature of Examinations for GCSE and A-level

1 GCSE

A student taking double subject science would be required to sit three 90 minute papers, for 80% of the total marks, and would also have a teacher assessment of coursework for the other 20%. For a single subject GCSE, say in physics, candidates would take one of the double subject's 90 minute papers plus a 60 minute physics extension paper, again with 20% coming from a teacher's coursework assessment. Single-award science would require three 60 minute papers for the 80%.

One typical 90 minute paper required 52 different responses, within parts of a set of structured questions, with all parts to be attempted. All responses are written on the exam paper and the number of lines for writing are set out in the paper. In the example described here, there were only four responses for which more than four lines were available for the response, the maximum being seven lines. Four of the 52 responses called for a calculation, all of which could be completed within three lines.

For coursework, a maximum of two pieces of work may be submitted, at least one of which must be 'a whole investigation'.

Comments

These very limited and uniform types of question follow from the view of the examiners that they must sample many different parts of the syllabus within the limited time permitted. They will also be influenced by the fact that short, highly specific demands are easier to mark in a uniform and defensible way: fear of variations between markers and of complaints and calls for re-marking may be a feature here.

A mixture of multiple choice and then a small number of more open-ended items might be a better use of examining time.

2 A-level

Edexcel Physics

The example here is Edexcel physics. The schedule is in two parts. Candidates have to take AS level at the end of their first year, and the results of this examination are combined with examinations at the end of their second year to yield the A-level result.

First year: AS level

Two 'unit tests' of 75 minutes each. Each contributes 15% of the A-level total.

Each paper comprises about eight structured questions, each in parts so that overall about 25 responses are required. All responses are written on the exam paper and the number of lines for writing are set out in the paper; in the example described here, there were only two responses for which more than four lines were available for the response. Nine of the 25 responses called for a calculation.

One practical test of 90 minutes, which contributes 10% of the A-level total.

A typical test requires work with three separate pieces of equipment, specified by the examining board. A fourth test question called for a plan for a specified investigation (i.e. no apparatus provided), with the response guided by a structure with about four separate parts. All responses are written on the exam paper and the number of lines for writing are set out in the paper; in the example described here, 18 separate responses were required, most of them requiring specified measurements or calculations or both.

One topics test of 30 minutes, which contributes 10% of the A-level total.

The paper consists of four questions, each on a separate topic; candidates have to answer one only (the topics are specified in the syllabus, so candidates only need to study one of them). All responses are written on the exam paper and the number of lines for writing are set out in the paper; in the example described here, each question called for between 11 and 15 separate responses, with about two having more than four lines available for the answer, and about three requiring any calculation.

Second year: components to be added to the AS results to determine A

One 'unit test' of 80 minutes, contributing 15% of the A-level total.

This is very similar in structure and style to the unit tests in AS described above.

One practical test of 90 minutes, contributing 15% of the A-level total.

This again is of the same structure and style as the AS practical.

One 'unit test' of 60 minutes, contributing 7.5% of the A-level total.

This is very similar in structure and style to the unit tests in AS described above, but having six questions requiring 18 different responses.

One 'synoptic test' of 120 minutes, contributing 20% to the A-level total.

This has four questions, all of which are to be attempted. Answers are to be given in a separate answer book, so the length of answers is not specified. Three of the questions are structured, each calling for four or five separate responses about a situation or experiment that is described in the question. The fourth presents a passage (about 400 words) describing a phenomenon (e.g. lightning), and asks for nine different responses to a set of structured questions.

Comment

It seems strange and regrettable that for every question the candidate is led by the hand through a set of specified steps. Studies of problem solving show that the most challenging aspect of a problem is to decide on which path to take, what sequence of steps to try, to cross the gap from the problem statement to its solution. Thus candidates, in the tests, and therefore in their preparation for the examination, are not faced with the real challenges of tackling problems.

There is no course-work component, and no requirement to conduct any practical exercise that might take longer than 15 minutes to complete, and for which step-by-step guidance is not provided.

Thus there is virtually no opportunity, and therefore no incentive, for pupils to write at any length about the science that they are studying, no incentive to devise a strategy for tackling a physics problem, and no incentive to spend time on designing and then performing an experimental investigation.

References

American Educational Research Association (AERA) with APA (American Psychological Association) and NCME (National Council on Measurement in Education) (1999) *Standards for educational and psychological testing* (Washington DC, AERA).

ARG (2002) T*esting, Motivation and Learning*; University of Cambridge School of Education: Assessment Reform Group.

Baker, E.L. & O'Neill H.F. (1994) Performance Assessment and Equity: a view from the USA. *Assessment in Education*, **1**, 11–26.

Ball, S.J. (1981) *Beachside Comprehensive*. Cambridge: Cambridge University Press.

Black, P.J. (1963) *Bulletin of the Institute of Physics and the Physical Society*, 202–203.

Black, P. (1993) Formative and Summative Assessment by Teachers. *Studies in Science Education*, **21**, 49–97. 1993.

Black, P. & Wiliam, D. (1998) Assessment and Classroom Learning. *Assessment in Education*, **5**(1), 7–71.

Black, P. J., Harrison, C., Lee, C., Marshall, B. & Wiliam, D. (2002) *Working inside the black box: Assessment for learning in the classroom*. London, UK: King's College London School of Education.

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

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

Bloom, B.S., Englehart, M.D., Furst, E.J., Hill, W.H. & Krathwohl, D.R. (1956) *Taxonomy of educational objectives. Handbook 1: Cognitive domain*, New York: David McKay.

Bransford, J.A., Brown, A. & Cocking, R. (1999) *How People Learn; Brain, Mind, Experience and School*. Washington D.C.: National Academy Press.

Butler, J. (1995) Teachers Judging Standards in Senior Science Subjects: Fifteen Years of the Queensland Experiment. *Studies in Science Education*.

Brown, M. (1998) Formative assessment for learning: general issues illustrated by examples from England. In Black, P. & Michel, A. (eds.) *Learning from Pupil Assessment : International comparisons. Centre for the Study of Evaluation Monograph Series No.12*. Los Angeles, CA: University of California, Los Angeles.

Cerini, B., Murray, I., & Reiss, M. (2003) *Student Review of the Science Curriculum*. London: NESTA.

Cohen, I.B. (1952) The education of the public in science. *Impact of Science on Society*, **3**, 67–101.

Cresswell, M.J. (1996) *Defining, Setting and Maintaining Standards in Curriculum-embedded Examinations: Judgemental and Statisitical Approaches*, chapter 5, pp.57–84 in Goldstein, H. and Lewis, T. (eds.) Assessment : Problems, Developments and Statistical Issues, Chichester and New York: John Wiley.

Daws, N. & Singh, B. (1996) Formative assessment; to what extent is its potential to enhance pupils' science being realised? *School Science Review*, **77**(281), 93–100.

DFES (2004) Interim Report of the Working Group on 14-19 Reform. Nottingham: DFES publications

Donnelly, J.F., Buchan, A., Jenkins, E., Laws, P. & Welford, G. (1996) *Investigations by order: policy, curriculum and science teachers' work under the national curriculum*. Nafferton: Studies in Science Education.

Donnelly, J.F. & Jenkins, E.W. (1999) *Science teaching in secondary schools under the national curriculum*. Leeds: Centre for Studies in Science and Mathematics Education, University of Leeds.

Driver, R., Squires, A., Rushworth, P. & Wood-Robinson, P. (1994) *Making sense of secondary science*. London: Routledge.

Driver, R., Leach, J., Millar, R. & Scott, P. (1996) *Young peoples' images of science*. Buckingham, UK: Open University Press.

Driver, R., Sduires, A., Rushworth, P. & Wood-Robinson, V. (1994) *Making sense of secondary science*. London: Routledge.

Dweck, C (2000) *Self Theories*. London; Taylor and Francis.

European Commission. (1995) *White paper on education and training: Teaching and learning – Towards the learning society*. Luxembourg: Office for Official Publications in European Countries.

Entwhistle, N.J. (1991) *Styles of Teaching and Learning*. Chichester: Wiley.

Fairbrother, R.W., Dillon, J. & Gill, P. (1995) Assessment at Key Stage 3: Teachers' attitudes and practices. *British Journal of Curriculum and Assessment* **5**(3), 25–31 and 46.

Gardner, J. & Cowan, P. (2000) *Testing the Test; a study of the reliability and validity of the Northern Ireland transfer procedure test in enabling the selection of pupils for grammar school places*. Belfast: Queen's University of Belfast.

Gardner, P.L. (1975) Attitudes to science. *Studies in Science Education*, **2**, 1–41.

Hacker, R.G. & Rowe, M.J. (1998) A longitudinal study of the effects of implementing a National Curriculum on classroom processes. *The Curriculum Journal*, **9**(1) 93–103.

Hacker, R.J. & Rowe, M.J. (1997) The impact of National Curriculum development on teaching and learning behaviours. *International Journal of Science Education*, **19**(9), 997–1004.

Harlen, W. & Deakin-Crick, R. (2003) Testing and Motivation for Learning. *Assessment in Education*, **10**, 169–208.

Harlen, W. (2004) A systematic review of the evidence of the reliability and validity of assessment by teachers for summative purposes. Research review produced for the EPPI Centre Institute of Education, London. In press.

Hashweh, M.Z. (1996) Toward and explanation of conceptual change. *European Journal of Science Education*, **8**, 229–249.
Hodson, D. (1998) The role of assessment in the curriculum cycle: a survey of science department practice. *Research in Science and Technology Education*, **4**(1), 7–17.
Hoge, R.D. & Coladarci, T. (1989) Teacher- based judgments of academic achievement: a review of literature. *Review of Educational Research*, **59** (3) 297 – 313.
Hoste, R.S. & Bloomfield, B. (1975) *Continuous assessment in the CSE: opinion and practice. Schools Council Examinations Bulletin 31*. London: Evans/Methuen Educational.
Hudson, P. & Smith, R. (1995) Teaching approaches in English science classrooms: has the National Curriculum really changed them ? Paper read at the San Francisco meeting of the American Educational Research Association.
Husén, T. & Postlethwaite, T.N. (1996) A brief history of the International Association for the Evaluation of Educational Achievement (IEA). *Assessment in Education*, **3**(2), 129–141.
Kellaghan,T., Madaus, G.F. & Airasian, P. (1982) *The Effects of Standardized Testing*. Boston: Kluwer–Nijhoff Publishing.
Linn, R.L (2000) Assessment and Accountability. *Educational Researcher*, **29**(2), 4–16.
Messick, S. (1989) Validity. pp. 12–103 in Linn, R.L. (ed.) *Educational Measurement (3rd Edition)*. London: Collier Macmillan.
Miller, R.W. (1989) *Fact and Method: Explanation, confirmation and reality in the natural and the social sciences*. Princeton, NJ: Princeton University Press.
Nuthall, G., & Alton-Lee, A. (1995) Assessing Classroom Learning : How Students Use Their knowledge and Experience to Answer Classroom Achievement Test Questions in Science and Social Studies. *American Educational Research Journal*, **32**(1), 185–223.
Osborne, J.F. & Collins, S. (2000) *Pupils' and parents' views of the school science curriculum*. London: King's College London.
Osborne, J.F., Simon, S. & Collins, S. (2003) Attitudes towards science: a review of the literature and its implications. *International Journal of Science Education*, **25**(9), 1049–1079.
Paechter, C. (1995) 'Doing the Best for Students': dilemmas and decisions in carrying out statutory assessment tasks. *Assessment in Education*, **2**, 39–52.
Rogosa, D. (1999) *How Accurate are the STAR National Percentile Rank Scores for Individual Students?–An Interpretive Guide*. CSE Technical Report 509a. Los Angeles, CA: CRESST. Published on web-site: http://www.cse.ucla.edu/products/reports_set.htm.
Russell, T.A., Qualter, A. & McGuigan, L. (1995) Reflections on the implementation of National Curriculum science policy for the 5–14 age range: findings and interpretations from a national evaluation study in England. *International Journal of Science Education*, **17**(4), 487–492.
Schibeci, R.A. (1984) Attitudes to science: an update. *Studies in Science Education*, **11**, 26–59.
Shulman, L. (1987) Knowledge and teaching: foundations of the new reform. *Harvard Educational Review*, **57**, 1–22.
Tobin, K. & Garnett, P. (1988) Exemplary practice in science classrooms. *Science Education*, **72**(2), 197–208.
Wandersee, J., Mintzes, J.J. & Novak, J. (1994) Research in alternative conceptions in science. In D. Gabel (ed.) Handbook of research in science teaching and learning (pp. 177–210). New York: Macmillan.
White, R.T. (1992) The origins of PEEL in R.J.Baird & I.J.Mitchell (eds.) *Improving the quality of teaching and learning: an Australian case study – the PEEL project*. Melbourne: Monash University.
Wiliam, D. (1996) Meanings and Consequences in Standard Setting. *Assessment in Education*, **3**(3), 287–307.
Wiliam, D. & Black, P. (1996) Meanings and Consequences: a basis for distinguishing formative and summative functions of assessment. *British Educational Research Journal*, **22**(5), 537–548.