NFER Position Paper on Assessment (2007)

Introduction

This paper has been produced in order to inform some of the current debates on National Curriculum Assessment in England. The Education and Skills Committee of the House of Commons has announced an inquiry into Testing and Assessment. In part this will examine testing and assessment in primary and secondary education as a key issue. Currently, pupils in England take Key Stage tests at 7 years-old, 11 years-old and 14 years-old in English, mathematics and science. This system has developed and evolved since its introduction in 1991. In January 2007 the Government announced that it would pilot several measures at Key Stages 2 and 3, including allowing pupils to sit national Curriculum assessments as soon as they are ready instead of waiting until the end of the Key Stage.

Our paper sets out the background to these debates, concentrating on the purposes of assessment, and the desirable characteristics which flow from these purposes. This leads to a statement of NFER's stance in relation to assessment. Finally, a commentary is given on two specific proposals for change currently under discussion: a national monitoring system based on a sampling approach; and the "Progress Tests" proposed in the DfES discussion paper "Making Good Progress."

Dimensions of Assessment

In essence, educational assessment serves two major purposes. The first of these is to provide immediate feedback to teachers and students, in order to facilitate the learning process. This is often termed formative assessment, but may also be referred to as diagnostic assessment. Recently it is also frequently referred to as *Assessment for Learning*. The second major purpose is to provide information which summarises a particular phase of learning or education. This information may be for a variety of institutional uses such as monitoring progress, certification or selection, and it may be for a variety of users, such as parents, teachers, governmental bodies and the learners themselves. This type of purpose is termed summative assessment. Both these purposes are important in the educational system as a whole, but they have different requirements and different characteristics.

This dimension of purposes is only one categorisation. A different categorisation, which cuts across the assessment purpose, is between formal and informal processes of assessment. The distinction here is between, on the one hand formal processes such as exams, tests and other assessments in which students encounter the same tasks in controlled and regulated conditions and, on the other hand, those less formal activities that form part of on-going teaching and learning. This second group would encompass question and answer, teacher observations, group discussion and practical activities together with classroom and homework writing and assignments.

Using this two-fold classification (as shown in the table below), it can be seen that formal and informal assessments can each be used for both formative and summative purposes. The formal processes are often managed externally to the school, though they need not be, and the informal processes are often internal to the school though they may provide information which is reported externally. The four cells of the table can be used to discuss the role and requirements of assessment systems and instruments.

	Purposes	
Processes	Formative	Summative
Informal	*Questioning* *Feedback* *Peer assessment* *Self assessment*	*Essays in uncontrolled conditions* *Portfolios* *Coursework* *NC teacher assessment*
Formal	*Analysis of tests, exams, essays* *Target setting*	*Tests* *Exams* *Essays in controlled conditions*

Formative Assessment

Informal Formative Assessment

Formative assessment is vital to teaching and learning. It is embedded in effective classroom practice and it should be tailored to the individual and their own stage of

learning. Such processes have long been regarded as essential for progress, providing a motivational effect for students as well as information on what has been recently achieved and the next steps in order to make progress.

Although such practices have always been intrinsic to successful teaching, recent research and policy has characterised them as a particular approach to assessment, leading to the principles that have been set out under the heading of "Assessment for Learning" and its characteristics are as follows:

- sharing learning goals with the pupils
- helping pupils know and recognise the standards they must aim for
- providing feedback that helps pupils to know how to improve
- both teachers and pupils reviewing pupils' performance and progress
- pupils learning self-assessment techniques to discover areas they need to improve
- pupils helping each other to learn
- including both motivation and self-esteem within effective assessment techniques.

Assessment for Learning is a simple idea which is far-reaching in its implications and quite difficult to put into practice. If teachers obtain information from assessment and use it to identify the next steps in learning, their teaching will be much more effective. Better still, if pupils are 'let in on the secret', so that they, too, understand what the next steps are, they will be better motivated and more successful learners. However putting this into practice well can make formidable demands on teachers in terms of their professional knowledge and skills.

We are very supportive of the principles of *Assessment for Learning* and believe these must be further promoted and new and better supportive materials must be produced and supplied to teachers. There is also a need for more, and more rigorous, research which explores the successful elements of *Assessment for Learning*. Our own work in this area has been concerned with providing helpful materials for teachers and in researching the possibilities of e-assessment in supporting *Assessment for Learning*.

There is a scarcity of good quality formative assessment materials to be used by teachers and students in classrooms. It seems to have been assumed that the very openness of formative assessment, and its devolving of responsibility to the student, renders such materials undesirable. In contrast, we believe that well-designed support materials can encourage the spread of formative assessment, and have undertaken projects to develop such materials, with the specific intention of fostering peer assessment in literacy for pupils. (See Twist, L. & Sainsbury, M. (2006) *Assessment for Learning: Literacy 10-11*. (London: nferNelson)).

It is often asserted that *Assessment for Learning* leads to greater gains in pupil's knowledge and understanding and these claims are impressive. We do though believe that there remains a need for more research evidence demonstrating what leads to such gains. There are limitations to *Assessment for Learning*, which arise from its classroom role. Because of its immediacy and the focus on what has just been learned and what is about to be learned, it is difficult to give information on the overall level of attainment or on the curriculum as a whole. The involvement of the teacher (and also the student as a self-assessor and other students as peer-assessors) introduces problems of reliability (and also bias) so that *Assessment for Learning* data is not necessarily good for comparing pupils.

A further difficulty is its detail. If it is to be used for summative purposes, then the essentially atomised data needed for *Assessment for Learning* alone needs collating in a systematic manner to allow an overall judgement which is reliable and comparable. This can be a time consuming task.

For all these reasons, we do not believe *Assessment for Learning* can or should provide summative information. Instead, we believe that this function should be a largely separate system with its own priorities, features and requirements.

Formal Formative Assessment

Formal assessments can also be used for formative purposes, whenever an analysis is made of performance in a formal test and the results used to plan teaching for classes, groups or individuals.

The national key stage tests over the last few years have been systematically analysed in order to draw out implications for teaching and learning, which have then been published on the QCA website; a large part of this work has been carried out by

NFER teams. An investigation of patterns of performance over a large sample of pupils can provide indications for teachers of typical patterns of errors. This can aid overall curriculum planning, but does not, in itself, give formative information for particular individuals or groups.

An additional problem with this approach is its timing. Formative information of this kind is of most use when the teacher is at the beginning of a programme of study, whereas the national tests are taken at the end of the key stage. A change in the timing of the national tests would in itself introduce greater potential for formative value.

A major focus of NFER's current work is the formative use of assessment information gained by more formal means. We are researching the potential of e-assessment in low-stakes contexts and to support assessment for learning. It is clear that teachers are required to focus on the understanding and attainment of individual pupils in order to develop effective plans for personalised learning. This will involve the management of a great deal of assessment evidence for planning teaching, in the form of test data and information on progress through the ongoing curriculum. E-assessment can occupy a central role, first in gathering detailed information about the nature of individual pupils' understanding and attainment, and then in collating and analysing this data. Rather than supplanting the teacher's role in relation to the child, it could supplement it, reducing the marking and recording workload while increasing and easing the flow of genuinely useful information.

In order to explore this opportunity an NFER research project is currently testing some of these principles. Experimental prototype questions are being trialled with samples of pupils and a variety of exploratory statistical analyses are being undertaken. This work may give rise to a clearer understanding of how e-assessment can provide a sensitive and unobtrusive evidence base for classroom activities and informative progress records.

Summative Assessment

Informal Summative Assessment

Since teachers have, by the very process of teaching, a wealth of informal assessment information on each pupil, there is a strong incentive to find ways of summarising that

information so that it serves a summative purpose. Ongoing informal assessment information covers pupils' performance across a range of contexts, and is thus potentially both more valid and more reliable than a single test result.

The National Curriculum assessment system recognises this by requiring teacher assessment judgements alongside test results. Although this is, and has always been, an intrinsic element of the system, it has tended to have been given less prominence than the test results. In the early 1990s, there were indications that the structured attainment targets and level descriptions were introducing a useful element of standardisation and a common language to teachers' informal assessments. This "community of practice" tended to decline around the beginning of the new century, however, because of the introduction of the national strategies, which had a strong focus on pedagogy but very little on informal assessment. However, over the last few years the balance has begun to change. The ideas of *Assessment for Learning* have been integrated into policy and with this has come a renewed interest in making use of informal assessment in more systematic and summative ways. Currently, the QCA initiatives on Assessing Pupils' Progress in secondary schools and Monitoring Children's Progress in the primary sector have reintroduced some of the original ideas and methods of the early National Curriculum, restructured in accordance with later thinking, technology and strategies. In Wales, the key stage tests have been replaced by a system of teacher assessment only, supported by publications in which standards are exemplified. NFER staff members have worked with DELLS[1] to develop optional assessment materials and exemplification to support summative teacher judgement.

In order to be used summatively, teachers' assessment information needs to be related to the standards which are provided by the National Curriculum level descriptions. However, the descriptions are broad and general, including many imprecise judgemental terms, so there is work to be done in reaching a shared interpretation of their meaning and application. This would involve a process of moderation between teachers, both within a school and between schools, which would require local leadership, possibly by a local authority adviser. Typically, the moderation process would involve discussion of specific pieces of pupils' work, chosen to represent the

[1] The Department for Education, Lifelong Learning, and Skills (DELLS) within the Welsh Assembly Government.

characteristics of a level, leading to agreement on the criteria to be applied. It would aim to result in an agreed, shared portfolio of exemplars. This process is professionally valuable but costly and extremely time-consuming.

A further time-consuming and potentially unmanageable aspect of informal summative assessment is the collection of evidence to support a judgement for each pupil. The system can collapse under an avalanche of paperwork if this is not managed carefully. The provision of an e-portfolio for each pupil could help greatly in managing the storage of examples of work and access to these, but will do nothing to reduce the time necessary to select, store, label and annotate the examples.

There is currently a debate about how far this kind of summative information can be used instead of test results, as in Wales and like coursework in public examinations. On the one hand, it has strong advantages in terms of scope and teacher involvement. On the other, its manageability is in question and its reliability has not been demonstrated.

Our view is that such a system in England is conceivable, but distant. There are three conditions that must be fulfilled before it could be introduced successfully. Firstly, a major investment – comparable to the introduction of the national strategies – has to be made in professional development in order to bring about a shared understanding of criteria. This would be supported by published exemplification materials and could include the use of some formal tests (as is currently the case at key stage 1). Secondly, a part of this professional development would need to address teachers' and advisers' understanding of the nature and purposes of the four quadrants of assessment, as described in this paper. It is necessary to reach a point where teachers perceive high-stakes summative assessment as professionally useful and complementary to formative approaches before a system of sufficient robustness could be introduced. Rigorous piloting and evaluation would be necessary in order to demonstrate appropriate levels of reliability. Finally, the system would need an element of external monitoring and accountability that commanded public and professional confidence.

Formal Summative Assessment

Formal summative assessment can serve many purposes. Among these are certification of schooling (as with GCSE) and selection (as with A-levels for university entrance). We will not consider these purposes here but concentrate on

summative assessment within schooling, principally through National Curriculum Assessment. This has had a consistent structure for about a decade, but there is currently renewed discussion on its purposes and methods. This has culminated in the department for Education and Skills' Consultation document "Making Good Progress" which proposes shorter, better focused "when ready" tests. This paper will give general observations on National Curriculum testing (for summative purposes) and specific comments on "Making Good Progress".

In commenting on testing in the National Curriculum the purposes of summative information need to be set out. Here, we are taking them to be, as follows:

A. The provision of comparable reliable information for children and their parents on their current levels of attainment.

B. The provision of comparable reliable information for children and their parents on the progress being made.

C. The provision of individual and grouped information for teachers to inform them of national standards and expectations in their subjects and to assist them generally with teaching pupils in the future.

D. The provision of grouped information for school managers and governors to inform them of the quality of learning of their students (and by inference the quality of teaching with the school) through the study of progress of their classes.

E. The provision of grouped school information for the public, providing an accountability function and contributing to choice for parents.

F. The provision of grouped school information to accountability agencies, such as LAs and OFSTED, to contribute to their judgements and measure improvement and decline.

G. The provision of central information to government and others on the education system as a whole, for monitoring standards over time and reporting on the curriculum in detail.

These seven purposes move from individual information to grouped information. They also move from levels of personal accountability to system accountability. It is a tenet of current government policy that accountability is a necessary part of publicly

provided systems. There is a broad consensus on this and we accept that accountability must be available within the education system and that the assessment system should provide it. However, the levels of accountability and the information to be provided are open to considerable variation of view. It is often the view taken of these issues which determines the nature of the assessment system advocated, rather than the technical quality of the assessments themselves.

It is worth remarking that in addition to the purposes set out above, National Curriculum tests have served other indirect but nevertheless important functions within the system.

H. For professional development of teachers, informing them of the nature of the National Curriculum and its interpretation. (This was particularly true of the early years of implementation, but continues to have a role. In some subjects, notably English, this has brought about a community of practice among teachers such that their judgements are much more aligned and standardised than they were before at Key Stages 1, 2 and 3. It is not necessarily also the case in mathematics, where many teachers continue to prefer test outcomes.)

I. To introduce positive change into the emphasis of the curriculum as taught (the delivered curriculum) – sometimes called a "backwash" effect[2]. Examples of this have included mental mathematics, spelling at key stages 1 and 2, and science processes at key stage 2 and 3.

J The accountability functions themselves contribute to a further indirect purpose for the assessment system, which has a political motivation: that of putting pressure on schools and teachers to maximise the attainment of pupils and students. The testing regime is intended to motivate students to perform to high standards, teachers to teach better and parents and school governors to raise the quality of schools. The underlying reason behind this is what is perceived as a stagnation in standards from the 1950s to 1980s at a time when educationalists alone were responsible for the curriculum and schooling. The rise of economic globalisation and the widespread belief that raising educational standards was vital to future economic survival, led to the

[2] The term "backwash is often used of the negative consequences of testing on the curriculum. The effects can though be either positive or negative.

accountability and pressure models of the current system. (Education of course, is not alone among public services in being subject to this sort of pressure.)

To these can be added some additional purposes which have arisen almost accidentally, but now have a useful function.

K In recent years there has been an acknowledgement of the importance of using national test data for school self-evaluation and improvement, often in partnership with other agencies such as the School Improvement Partner. The provision of sophisticated indicators based on national testing data, such as DfES/Ofsted's Contextualised Value Added (CVA) measures or those provided by the Fischer Family Trust (FFT) has led to a significant improvement to schools' ability to evaluate their own performance. These indicators rely crucially on the current national testing system, and any replacement system proposed would need to offer equivalent or better measures if there is any desire not to lose the progress which has been made in this area.

L The availability of comprehensive national data with attached to it detailed pupil information has provided a powerful tool for the evaluation of the impact of educational initiatives on attainment and performance. Examples include NFER's work on evaluating Excellence in Cities, the National Healthy School Standard, Playing for Success, and the Young Apprentices Programme. Such data provides an important instrument for informing educational policy.

This account of the purposes of National Curriculum Assessment shows that there are many of these, and any calls for change needs to consider which are the most important and which can be downgraded. In the existing system, the current National Curriculum tests are a compromise which attempts to meet all these purposes. The accountability functions mean that they must achieve high levels of reliability. This means that the results must be reliable and subject to a limited amount of error and misclassification. (It is important to recognise that all tests, indeed all judgement processes have some component of error – this includes examinations, interviews, teacher judgement, and legal processes.) Any development of the existing system and its tests for which the accountability purposes remain, would properly need to

demonstrate that it has equivalent or higher reliability. We do not believe it would be defensible to have a system in which levels of reliability are not known or cannot be demonstrated.

As one of the developers of National Curriculum tests, we are aware of the thorough development process they undergo and the underlying statistical data on their performance. In our view, the current tests achieve the necessary technical and psychometric requirements to a reasonable extent. They have good to high levels of internal consistency (a measure of reliability) and parallel form reliability (the correlation between two tests). Some aspects are less reliable, such as the marking of writing, where there are many appeals / reviews. However, even here the levels of marker reliability are as high as those achieved in any other written tests where extended writing is judged by human (or computer) grades. The reliability of the writing tests could be increased but only by reducing their validity. This type of trade off is common in assessment systems with validity, reliability and manageability all in tension.

The present tests do provide as reliable a measurement of individuals as is possible in a limited amount of testing time. When results are aggregated over larger groups such as (reasonably large) classes or schools, the level of reliability is extremely high.

A second requirement of the National Curriculum tests (and all assessments) is that they should be valid for their purpose. According to current thinking[3], the validation of a test consists of a systematic investigation of the claims that are being made for it. In the case of National Curriculum tests, the claims are that the tests give an accurate and useful indication of students' English, science or mathematical attainment in terms of National Curriculum levels. The tests do have limited coverage of the total curriculum: the English tests omit Speaking and Listening, the science tests formally omit the attainment target dealing with scientific enquiry (though questions utilising aspects of this are included) and mathematics formally omits using and applying mathematics. Outside of these the coverage of content is good. The fact that the tests

[3] See, for example: American Educational Research Association, American Psychological Association and National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing.* Washington, DC: AERA.

change each year means that the content is varied and differing aspects occur each year. In general, the content validity of the tests can be regarded as reasonably good in relation to this coverage of the National Curriculum. However, a full validation has other aspects and these are seldom considered in relation to the National Curriculum tests, principally because of their numerous purposes. In general, the current tests adequately serve the accountability requirements, listed above as A to F. They may not meet the monitoring requirement (Purpose G) so well and we address that below.

We therefore believe that there should not be changes to the existing system without careful consideration of what the purposes of the system are and a statement of this. Any proposals for change should set out carefully which of the above purposes they are attempting to meet and which they are not. The level of requirements for validity and reliability could then be elucidated and the balance with manageability and the resources required determined. If accountability is no longer to be required then a different assessment regime could be implemented. However, this should not be done without evidence that any replacement would meet its own purposes validly.

NFER Assessment Philosophy

The NFER view of assessment is to acknowledge and embrace the variety of assessment purposes and processes that the discussion above has set out. Both broad purposes and both types of process have their place in the overall assessment enterprise. It is meaningless and unhelpful to dismiss summative assessment because it is not formative, or to dismiss informal assessment because it is not formal. Our work encompasses all four quadrants and it is important to recognise the distinctive features and requirements of each.

Correspondingly, the need is for education professionals and policymakers to develop the same kind of understanding. The classroom teacher, like the assessment researcher, is required to deal with all four quadrants. The best approach to this is to understand and accept the distinctions and relationships between them, and to give appropriate attention to each one. Similarly, policymakers, officials and teacher educators must recognise that teachers have this variety of assessment responsibilities and opportunities and give attention and respect to all of them.

Our stance in relation to assessment is that there must be a clear statement of the intended purposes of the assessment system and that its processes and instruments should have an appropriate level of validity and reliability to provide sound evidence for those purposes. This implies that there should be a sound development process for instruments, and evaluative research to demonstrate that the judgements being reached on the basis of the system are soundly based.

Specific Proposals for Change

National Monitoring

One of the current purposes of National Curriculum Assessment is the provision of central information on the education system as a whole, for monitoring standards over time and reporting on the curriculum in detail (purpose G). It is here that the present system may be less valid. First, there are difficulties in maintaining a constant standard for the award of a level in a high stakes system where tests or questions cannot be repeated. We do though believe that the methods used for this currently which include year-on-year equating and the use of a constant reference point through an unchanging "anchor test" are the best available and lead to the application of a consistent standard. A second consideration is that the curriculum coverage each year is limited to the content of that year's tests. In response to these (and also other issues), there has been considerable advocacy of a light sampling model for monitoring the curriculum and changes in performance.

National Curriculum Assessment currently has monitoring national performance as only one of its many purposes, and is probably not optimal for this, as is the case for most assessment systems which attempt to meet many purposes. NFER conducted a review of educational statistics across the UK for the Statistics Commission, which was included in their report on the subject[4]. They concluded that the current national monitoring system in England was sufficiently fit for purpose that an additional survey would not be cost-effective.

We believe that, in principle, if the sole goal of an assessment system is to derive comparable measures of national attainment at different time points, then a low-

[4] See: http://www.nfer.ac.uk/publications/other-publications/downloadable-reports/pdf_docs/serfinal.PDF

stakes, lightly-sampled survey is probably the best way of meeting this one aim. Low-stakes testing has the advantage that there is no incentive to 'teach to the test', reducing the effects in schools. (Though, as we have seen a positive backwash effect is one of the current uses of National Curriculum tests.) Because of reduced or negligible security issues it is possible to repeat substantial numbers of items from survey to survey, thus enabling relatively reliable measures of change over time to be adduced. It may not be necessary to monitor national performance on a yearly basis, and in this case less frequent surveys would be possible. A well-stratified national sample should enable good estimates of the uncertainty in the national performance measures to be made. A matrix sampling design, in which different pupils take different combinations of test items, would enable a wide coverage of curriculum areas to be maintained while minimising the burden on individual pupils.

However, there are some problems with this approach, which should be recognised. The lightly-sampled low stakes assessment would provide one view of standards, but because it is low stakes it may well underestimate what students are really capable of when they are more motivated. Our experience and the research literature shows that there is a large difference in scores on the same test in high and low stakes situations. This is a validity issue related to the difference between performance in motivated and unmotivated conditions. If we are interested in monitoring what pupils can achieve when not under motivated to achieve, low stakes surveys are well and good. If we are interested in performance when the results matter, this approach would not give it. It would also mean that such survey results would not align with any high stakes measures that continue e.g. GCSE.

There is considerable opposition in schools to taking part in optional assessment exercises, particularly secondary schools. However, anything other than a very high school response rate would cast serious doubts on the results, due to non-response bias, but it would be hard to find suitable incentives for schools to take part. Problems with response rates in international studies such as PISA, TIMSS etc. illustrate this – considerable efforts have been put into the attempt to persuade enough schools to take part to achieve the sample response rate constraints. It would probably be necessary in the modern climate to make participation in the survey compulsory for the selected schools in order to assure proper representative national samples.

Nevertheless, we would support the introduction of a properly planned regular national monitoring exercise, to examine changes in performance at regular intervals, on a sample basis, and to monitor the curriculum widely. To assess the full curriculum in a valid manner may well require assessment methods other than written tests (e.g. for speaking and listening, science experimentation). Such methods were attempted in the Assessment of Performance Unit (APU) surveys in the 1980s, conducted by the NFER and others, but proved difficult and expensive to implement. The lessons of the experience of that monitoring exercise also need to be learned. It would need to be regarded as a proper research exercise with the collection of background data on pupils and schools, in order to examine educational and social questions. It should also ensure a wide agreement on the appropriateness of its methodology and analysis techniques, reducing the possibility of attacks on its results.

Making Good Progress Proposals

The Making Good Progress proposals range widely over assessment, personalised learning and target setting. These should properly be regarded as a whole. However, since this paper deals with assessment issues, we will concentrate on that part of the proposals. Within the Making Good Progress document, it is proposed that there should be a new type of tests. The features of these are described briefly, and appear to be as follows:

- Single level tests
- Testing when ready – shorter more focused and more appropriate tests
- Externally set and marked
- "One way Ratchet" - never going back, only forward

In general, we are supportive of the notions of testing when ready and the close tie to teaching and learning. This fits within the context of Personalised Learning/ Assessment for Learning. As such, "progress tests" could provide a useful stimulus to teaching and learning. However, as described in Making Good Progress, we would doubt that they can fulfil that function. As a single level test, awarding a level, the test would generally show what a student could do but it would not be able at the same time to provide diagnostic information about the next steps since these would

not be included in the test. Similarly, because it would have to cover the curriculum broadly at that level, and levels represent two years of teaching (on average), it could not identify the small next steps needed for personalised learning.

For these reasons, we do not believe that the tests as described could support teaching in any direct way. If this is the desired intention, a different model with a suite of short tests, relating to specific elements of the curriculum, and providing information both on what has been achieved and the next steps would be more appropriate. There would need to be a large bank of such tests available for testing when ready on an individual basis. To be most useful they would be marked by the teacher, immediately, rather than through an external system. Such tests would be low stakes and have little accountability function.

In fact, Making Good Progress makes clear that the proposed progress tests would be used for accountability purposes, with the levels awarded being retained and reported. This means that the tests will need to have the characteristics of tests for accountability: high levels of reliability and validity. The following sections examine the proposals from this viewpoint.

The meaning of the phrase 'single level tests' will need some exploration. In a sense, the existing tests (or tests in the same style) could be utilized as tests which simply give a pass/fail at a single level. Given their length and their coverage of the curriculum this would lead to results with a reasonable (and measurable) level of reliability. However Making Good Progress states that the tests will be 'shorter and more focused.' There is a strong relationship between reliability and test length, so an unfortunate implication of this is that the tests will have lower levels of reliability and reduced curriculum coverage.

In this context, the important aspect of reliability is the consistency of the decisions made. It there were two progress tests at the same level, what would be the percentage of students classified the same way on both occasions? For the tests to be shown to be useful, this needs to be considerably above chance levels. In the current reading and writing tests at key stage 2, the degree of decision consistency for each level is at least 80 per cent and for some levels is as high as 98 per cent. The progress tests would have to match these levels of consistency. This would need careful examination

during development, as reducing the length of test inevitably leads to lower levels of reliability.

A second aspect of the 'shorter more focused' approach is curriculum coverage. In the current National Curriculum tests considerable efforts are made to include as wide a representation of the curriculum as is practically possible in a written test. This is essential for demonstrations of validity. Moreover, the annually changing tests mean that, over time the tests have even wider coverage. In writing for example, different text types/genres are sought from children each year and, within the test each year, two different tasks are required. Hence, reducing the length of the tests could also reduce the validity of the test.

The concept of testing when ready can be a useful one, particularly if it is used formatively and incorporated into the teaching-learning process as in *Assessment for Learning*. However, its utility within a summative system may not be as apparent. The provision of information from the "progress tests" (which are aimed at making judgements about a single level) is unlikely to have the diagnostic element useful for *Assessment for Learning*. The argument advanced in Making Good Progress is that success at one level will stimulate progress toward the next level, acting motivationally. This will need to be evaluated in practice. It may be that the levels are so far apart (they are intended to cover two years of development) that achieving one may actually slow progress, since the next target may be too distant. This is particularly a concern because of the 'one way ratchet' proposal. The achievement of a level and the knowledge that it cannot be removed may act to demotivate rather than motivate.

We have further concerns about the 'one way ratchet'. Its underlying assumption seems to be that children's learning is an ordered progression and that movement is always forward. This is not in fact the case, and children can decline in terms of skills or knowledge. It is therefore useful to have later checks that a level previously achieved has been maintained. If this is not the case, we do not believe the "one way ratchet" should be implemented.

This issue may interact with that of the reliability of the test. If the decision consistency of the tests at a given level is low, then a large proportion of candidates could be misclassified as achieving the level when they should not. If this is coupled

with the ‘one way ratchet’, the misclassification would become enshrined, possibly being harmful to such children’s progress as they would be being treated (and taught) as if they were at a higher level than was actually the case.

It is not the case that the levels of the National Curriculum are, in practice, as even and well ordered as the underlying model would suggest. In a given strand of a subject, the difficulty of the content may not increase in regular steps. Similarly, in different strands, the difficulty of the processes or skills at a given level may not be the same. It was this type of difficulty that led to the abandonment of the strong criterion referencing model of the early national curriculum assessment in the 1990s. This was replaced by a weak criterion referencing model in which content from various levels and across a broad range has been included in the National Curriculum tests, leading to the setting of an overall subject level (or within English, reading and writing levels). It also marked a return to a traditional psychometric principles and a mark-based scoring system.

There is a naïve view that questions can be written at a single level, derived from the level descriptors and these will have comparable difficulty. Taken to its extreme, it is sometimes thought that a single level test could be constructed by having material drawn from the level descriptor at that level. Candidates would then be expected to answer a set proportion of this correctly. This might be 50 per cent, or more usually 80 per cent, or sometimes all. There have been examples of such systems which have been constructed with these principles and in which the consequence has been very low pass rates. We therefore would advise that although the Progress Test may award a single level, they should have the following characteristics:

- Sufficiently long (in terms of numbers of questions and marks awarded) to have a good curriculum coverage, leading to good evidence of validity.
- Sufficiently long (in terms of numbers of questions and marks awarded) to have high levels of reliability, so that decision consistency is good and the number of misclassifications (particularly false positives) is small
- Include content from the level below as well as the target level in order to elicit a range of outcomes, and also to allow some simple questions to give pupils confidence and to motivate them.

- Include content from the level above as well as the target level in order to elicit a range of outcomes, and to allow some formative information to be provided for next steps.
- For writing, continue to allow a range of levels to be demonstrated, through differentiation by outcome.
- Set the criterion for achieving the level through soundly based equating or judgemental processes, not through the application of strict algorithms which assume equal difficulty in questions and in tests.

In addition, we would advise that the 'one way ratchet' is abandoned and that the system allows for re-testing of doubtful cases so that high levels of certainty are achieved and so that misclassification is minimised. A useful refinement would be to have a system in which there are three levels of outcome: level X awarded; level X not awarded; and a band of uncertainty in which a retest is advised in the following test round. Hence teachers could report only success which is assured to a high probability, requiring pupils with scores in a defined range of uncertainty to be retested.

To summarise, we do believe that some version of Progress Tests may be a useful addition to the system, but believe that their purpose must be carefully defined. That purpose should then lead to a specification and a development process that produces tests which are fit for use in terms of their reliability and validity.