Educational Research
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713699076

Considering alternatives to national assessment arrangements in England: possibilities and opportunities

Sylvia Green [a]; Tim Oates [a]
[a] Cambridge Assessment, Cambridge, UK

To cite this Article Green, Sylvia and Oates, Tim(2009) 'Considering alternatives to national assessment arrangements in England: possibilities and opportunities', Educational Research, 51: 2, 229 — 245
To link to this Article: DOI: 10.1080/00131880902891503
URL: http://dx.doi.org/10.1080/00131880902891503

Educational Research
Vol. 51, No. 2, June 2009, 229–245

Routledge
Taylor & Francis Group

Considering alternatives to national assessment arrangements in England: possibilities and opportunities

Sylvia Green* and Tim Oates

Cambridge Assessment, Cambridge, UK

(*Received 11 July 2008; final version received 2 February 2009*)

Background: In this article we address some of the challenges posed by the development of national assessment systems and discuss the need for high quality information on trends in attainment; support for school improvement processes and ways in which learning should be enhanced through valid assessment.
Purpose: Key elements are explored, including the need to: monitor national standards; provide accountability data; give feedback to learners, teachers and parents. The dangers of multi-purpose testing are outlined in the context of political and educational objectives and the need to separate out the functions in an educationally valid way are considered.
Evidence: Experiences in England are used to illustrate some of the key issues and problems that need to be taken into account when designing effective national assessment models. We refer to the national debate and policy reviews on assessment issues. We also draw on a body of literature related to this field of knowledge.
Main argument: A range of possible models of assessment are outlined and their advantages and disadvantages discussed. The aim is to offer potential systems for consideration in an attempt to promote the design and development of assessment processes that will generate valid and reliable data on attainment for individuals, schools and at a national level.
Conclusion: The challenge is to find ways of achieving these objectives within a framework of educational quality through the enhancement of teaching, learning and assessment.

Keywords: national assessment; purposes of assessment; accountability; national standards; formative assessment; management of change

Introduction

The development of models of national assessment poses many challenges and it is important to consider such development in the context of the need for high quality information on trends in attainment; support for school improvement processes and ways in which learning should be enhanced through valid assessment. In this article, we explore a range of key elements including:

- monitoring national standards;
- school accountability;
- feedback to learners, teachers and parents.

*Corresponding author. Email: green.s@cambridgeassessment.org.uk

ISSN 0013-1881 print/ISSN 1469-5847 online

DOI: 10.1080/00131880902891503
http://www.informaworld.com

This article is not suggesting that all systems have these three functions at their heart; the analysis suggests that they are the key functions in the English system at the current time, with the accountability agenda being of increasing interest in nations anxious to enhance the quality of schooling and the performance of educational systems (Wößmann et al. 2007). Not all systems seek all three of these; this article gives models and insight into contexts in which all of these are the key objectives. It also outlines the dangers of multi-purpose testing in the context of political and educational objectives (Newton 2007). Educational values and validity are discussed alongside the importance of 'fitness for purpose' in systems that promote and enhance teaching and learning. Experiences in England are used to illustrate some of the key issues that need to be taken into account when designing effective national assessment models and the factors that militate against real change. The article concludes by outlining a range of possible models of national assessment in an attempt to promote the design and development of assessment processes that will generate valid and reliable data on attainment for individuals, schools and at a national level. The apparent challenge is to find ways of achieving these objectives within a framework of educational quality through the enhancement of teaching learning and assessment.

Curriculum and assessment development

The current assessment system in England (statutory pupil assessment, including an element of teacher assessment but operating principally through external national tests in mathematics, English and science, at the end of each key stage of education, typically ages seven, 11 and 14), although subject to constant amendment, has essentially been in place since the early 1990s following a report on assessment and testing in 1987 from the Task Group on Assessment and Testing (TGAT 1988), set up by the government of the day and chaired by Professor Paul Black, King's College, University of London. Two decades later, the Select Committee Inquiry on Assessment and Testing (House of Commons Children, Schools and Families Committee 2008) re-ignited more detailed debate in UK government preceding an unexpected, immediate cessation of national testing at key stage 3 (principally 14 year olds) in October of the same year, following a breakdown in test marking (Sutherland 2008). The subsequent discussions of potential alternatives reinforced the importance of understanding the relationship between assessment and impacts on learning, such that any new model of national assessment is designed to enhance that relationship (Cambridge Assessment, National Foundation for Educational Research and the Nuffield Foundation 2009).

One of the key issues emerging from these debates is that the warnings about multi-purpose testing have not been heeded, and that the influence of assessment on curriculum and on institutional behaviour and strategy escalated (Mansell 2007; Wiliam 2003; Wellcome Trust 2008). But the issue of conflict between, and balance of, different purposes was hardly a new consideration. One of the proposals of the TGAT report (1988) was that assessment:

> ... should be an integral part of the educational process, continually providing both 'feedback' and 'feedforward'. It therefore needs to be incorporated systematically into teaching strategies and practices at all levels. Since the results of assessment can serve a number of different purposes, these purposes have to be kept in mind when the arrangements for assessment are designed. (I.4)

On publication of results, the report highlighted that:

> ... there is a fear that results will be published in league tables of scores, leading to ill-informed and unfair comparisons between schools. We believe that most teachers and schools would not object to assessment results being reported to those who know the school and can interpret them in the light of a broader picture of its work and circumstances. (III.18)

And went on to recommend that:

> ... the only form in which results of national assessment for, and identifying, a given school should be published is as part of a broader report by that school of its work as a whole. (XII.132)

The vision of TGAT was that the national assessment system should be essentially formative, including provision of information to allow parents to understand the rate of their children's progress through the national curriculum. Summative purposes would be focussed on assessment at age 16. It was also envisaged that standardised assessment instruments including tests, practical tasks and observations would be used in order to minimise curriculum distortion and that the system would be based on a combination of moderated teachers' ratings and standardised assessment tasks. National assessments were to take place at ages seven, 11 and 14 years, at the end of the phases of education known as key stages (TGAT 1988; Sutherland 2008).

One of the aims of the assessment system was to minimise negative wash-back into teaching and learning by including valid tasks that would encourage effective pedagogy and would assess the parts of the curriculum that paper and pencil tests could not reach (TGAT 1988).

The standards agenda

It is important to recognise that the introduction of a national curriculum and a national assessment system in England led to a number of positive outcomes. Certain forms of national coherence improved (Chitty 1999; Ofsted 2002; Thomas et al. 1997), expectations were increased through a challenging curriculum (Barber and Graham 1993; Hopkins and Sebba 1995) and certain categories of learners derived particular benefit in terms of progression and attainment (Dobson and Henthorne 1999; Arnott 2004).

The historical record is sparse in respect of early decisions around the functions of national testing. What is clear is that sample-based monitoring of national standards, via the Assessment of Performance Unit (Gipps and Goldstein 1983; Newton 2008), was brought to a close, in the face of national datasets on individual pupil attainment at ages 7, 11, and 14 and more elaborated national data from public examinations. Monitoring national attainment standards joined the formative and evaluative functions of national testing (TGAT 1988). Improvements in national test outcomes were thus very strongly associated with putative gains in national attainment (Clarke 2003). Two related questions thus emerged: to what extent did the changes and initiatives introduced at that time lead to the gains in test performance that have been so widely celebrated and to what extent could the national test results be relied upon as an infallible indicator of standards of

attainment over time? A project was commissioned by the then regulatory authority, the Qualifications and Curriculum Authority (QCA) in England, to investigate the comparability of test standards and was carried out by researchers at Cambridge Assessment (Massey et al. 2003; Massey 2005).

The stability of test standards at ages 7, 11 and 14 years of age in English, maths and science from 1996 to 2001 were investigated. The findings were varied across age groups and subjects with some tests appearing more lenient and some more severe over time. However, the evidence also suggested significant gains in achievement even where national test results may have exaggerated their extent. The question of maintaining test standards over time is central to any discussion of improved performance and is a problem facing any country introducing a national assessment model to measure standards and changes over time.

There are a number of other negative impacts of national testing, which are well rehearsed in the literature. There is evidence to suggest that increases in performance are often found when high stakes tests are introduced because teachers and students become familiar with the test requirements rather than as a result of real improvements in learning (Wiliam 2001, 2008). Negative effects occur when too much time is spent on memorisation, question spotting and test practice, to the detriment of positive teaching and learning (James 2006; Mansell 2007). Anxiety is also an issue as is student de-motivation with added pressures from league tables and target setting. In their review of research *Testing, motivation and learning* (ARG 2002), the Assessment Reform Group, an influential group of researchers in assessment, found strong evidence of the negative impact of testing on pupils' motivation. For the less able, lack of success was found to lower self-esteem, leading to an increased gap between high and low achievers. Rising instrumentalism in learners and orientation to surface learning in pedagogic strategies amongst teachers has also been detected (Black and Wiliam 1998; Broadfoot et al. 1998; James and Gipps 1998) and has been accompanied by increased incidence of 'maladminstration' of national testing in the period 2001–07 (Hansard 2008). Alongside criticism of the consequences that have flowed from the national assessment model in England, fundamental questions have been raised regarding 'fitness for purpose' (Wiliam 2001; House of Commons Children, Schools and Families Committee 2008; Oxfam 2008). Indeed, by February 2007, these accumulating problems led UK government to trial a revised model – Single Level Tests (SLTs) – as a potential mechanism for refining the overall assessment system.

Until the announcement of the SLT trial, there had been a history of minor modifications and refinements, which mainly involved 'bolting on' additional requirements, for example, in relation to optional tests and additional national data collection. However, one of the major problems with trying to implement fundamental change is that the national test data are used for many purposes and by several agencies, for funding decisions, inspection evidence, league tables etc. The intertwined nature of different administrative and operational functions and agencies in the system introduces substantial resistance to change to the system. This can be described as 'interlinked dependent complexity' where multiple functions and operations depend on the assumptions of national testing, and the precise data being the 'food and fuel' of so many parts of the system.

As Oates (2008) emphasises, significant impetus is required to escape the gravitational pull of existing arrangements. He suggests an illustrative metaphor. You can launch a projectile into space on its way to new planets, but if it has

inadequate energy, it will fall back to earth, and you will end up near where you started, albeit at great expense and with quite a lot of wreckage. This captures the process that occurred throughout the 1990s and into the twenty-first century in respect of national testing in England; new developments seemed to lack the escape velocity to ensure that their purpose, form and operation were genuinely progressive. Innovations were dragged back, by the pull of existing culture, opinion and processes, to a position that reproduced existing arrangements.

The first adopted alternative model of national assessment, SLTs, was launched by its civil servant authors, in early 2008, as a putatively radical development of national test arrangements (National Assessment Authority 2008). The aim of this development was reported to be to create 'when ready' tests, which could be taken by students at a time deemed appropriate by the teacher rather than as in current arrangements where national assessments are administered at ages seven, 11 and 14 (National Assessment Authority 2008).

These 'new' tests, as their name suggests, assess at one level of the National Curriculum rather than the existing tests, which typically address three levels. Responses from assessment experts during the consultation, which followed the launch of the pilot for the tests, suggested that this new model was insufficiently distinctive from current arrangements, and that a range of fundamental measurement issues would prove troublesome in the piloting and operation of the tests (NFER 2007; Cambridge Assessment 2007; MEI n.d.).

In the first administration of the tests (December 2008), the majority of the problems were indeed realised. Announcements have now been made (March 2008) regarding a shift in emphasis from using the tests to confirm that learners are 'secure' in a national curriculum level to 'threshold performance' in a level, i.e. back to the current focus. There has also been some exploration of the possibility that the 'new' tests might cover more than one level (BBC News Online 2008). If these changes are implemented, the supposed radical features of the new arrangements are to be diluted, and the testing arrangements will be far closer to simply providing two sessions, per year, of the existing test model. This brings the risk of testing further dominating the school curriculum (Mansell 2007) – hardly the intended effect of the original innovation. The experience of the SLTs so far brings to mind the sentiment of H.L. Mencken: 'For every complex problem there is an answer that is clear, simple and wrong.'

In developing any new model the degree of complexity needs to be genuinely addressed – including the interrelatedness of different functions and agencies in the system, unintended consequences and adjuvant policy designed to secure beneficial wash-back from assessment. Crucial to this is separation of the functions of the assessments and recognition of the specific risks of using assessment instruments simultaneously for assessing the individual student and inform learning, for national monitoring and to call teachers and schools to account.

Many have argued that the publication of national league tables in England and the pressure this places on teachers, schools and students has had a detrimental effect on teaching and learning; essentially because the accountability function impedes the ability to use assessment as an integral part of the learning process, and placing the teacher in a conflicted position (James 2006; Mansell 2007; Stobart 2008). Part of the explanation for the development of this set of circumstances derives from the fact that there has been an increasing interest in the detail of performance, both on the part of UK government (in respect of school accountability and national

standards) and on the part of teachers and school managers (feedback on strengths and weaknesses in provision). Whilst systems increasingly are able to generate fine-grained information and systems for storing and displaying it are becoming increasingly elaborated, a key question emerges: are we matching our development of such systems with processes by which we can make valid inferences on the basis of these data? There is a potential problem when teachers do not have the skills or techniques to handle a complex array of data, and are not yet able to use the data as a basis for differentiated, 'personalised' learning to any great extent or, indeed, with adequate validity.

Measuring performance of national education systems

Measurement of the performance of national education systems has become an increasing matter of interest both for national governments and for transnational organisations conducting international surveys (Monseur, Sibberns and Hastedt n.d.; Postlethwaite n.d.). Issues such as the effects of innovation and change, the differential performance of different groups in society and trends in standards over time are all key issues in monitoring and managing national systems.

The OECD's Programme for International Student Achievement (PISA) has joined the IEA's Trends in Maths and Science Study (TIMSS) and Progress in Reading and Literacy Study (PIRLS) as pre-eminent international studies allowing nations to both reflect on their own performance and compare their performance with that of others (Ofqual 2008).

However, such studies are not unproblematic as instruments for national governments interested in close and accurate scrutiny of the performance of their own systems (Oates 2007). Although wide-ranging, these studies are not undertaken at optimum times for national systems, e.g. nations may be introducing innovations that require system-level monitoring (Cambridge Assessment, National Foundation for Educational Research and the Nuffield Foundation 2009). Other problems include: such studies are insufficiently frequent for national monitoring purposes; the content changes over time and they thus have problems in being valid measures of change in student attainment over time; they can suffer from 'low stakes' and sampling problems; and they are not highly sensitive to the specific curriculum and assessment arrangements in specific national settings (Oates 2007; Ofqual 2008).

Nation-specific processes of system monitoring persist, even in the presence of ambitious studies such as PISA. Most notable of these are the National Assessment of Educational Progress (NAEP) in the USA, the New Zealand NEMP (National Education Monitoring Project) and the Scottish Survey of Achievement and – very different in form from these – English national curriculum assessment and Performance Tables.

The current American NAEP, the NZ NEMP, the Scottish survey and the now-redundant English APU closely resemble one another in purpose and shape. They involve independent, low-stakes tests, which maintain consistent content over time and are used to assess a representative sample of children, with a matrix sampling method being used to cover the full range of curriculum content. Such systems benefit from stability in measures (allowing robust measurement of standards over reasonable timeframes), fuller coverage of the curriculum, lack of distortion deriving from 'teaching to the test' and comparatively low cost. They suffer from problems of declining relevance of content, absence of direct motivational wash-back into schools

and student performance, and failure to be valid measures of performance at school level. They do not provide feedback to parents on each and every child, nor do they link with national examinations/tests or teachers' assessments of their own students in the classroom. Finally – and crucially – it is naïve to think that they are free of the typical range of operational and contextual pressures which affect large-scale technical exercises in national education and training systems (Oates 2007; Ofqual 2008). Indeed, the APU was severely compromised by resourcing constraints, methodological disagreements, and imposed, restrictive deadlines (Gipps and Goldstein 1983; Oates 2007) – although its ultimate demise derived from crude assumptions, at national policy level, that national assessment – giving information every year on every child reaching the end of key stages 1, 2 and 3 – would be an entirely adequate substitute for matrix-based sampling.

The National Curriculum and Performance Table model in England relies on using data from each and every child (from national assessments undertaken at 7, 11 and 14, and national subject-based examinations taken at 16 and 18) to build a national picture of educational attainment standards, to provide information for learners, parents and teachers and to judge the performance of schools. At national policy level, the availability of data on each and every child has been seen at the bedrock of accountability systems, the data being used as the key system management tool within public policy.

But, through problematic conflation of a variety of purposes (Newton 2007) and underlying measurement issues (Massey et al. 2003), major structural problems accumulate not merely in the form of undesirable 'wash-back' into the curriculum (such as the aforementioned 'teaching to the test' rather than focusing on 'deep learning' (ARG 2002; Mansell 2007), but also in respect of necessary annual change in the content of 'high stakes' tests and examinations, in order to safeguard security; failure to cover full curriculum content in each test/exam session; and misclassification of attainment in terms of 'levels' (Wiliam 2001). The extent to which ALL purposes are compromised by a conflation of purposes is a key issue in the English context (Newton 2007; Oates 2007). We would like to suggest that separation of purposes and careful alignment of these with adequate and well-matched operational arrangements to deliver on these purposes is vital in respect of responsible and efficient public policy. To this end, in the light of the fundamental aims of national assessment in England, we suggest that the principal functions of the assessment systems and allied arrangements should be to:

- deliver information to pupils, parents and teachers to enhance learning;
- operate systems of accountability for schools;
- deliver highly robust information on system performance, for policy purposes.

The second purpose remains controversial. While accountability arrangements present a vital linking between democratically derived aims for education and training and delivery systems, the focus (in terms of which 'level' in the system should be the 'unit of interest' for enhancing attainment (Cambridge Assessment, National Foundation for Educational Research, and the Nuffield Foundation 2009) and the form of arrangements remain contested. As a means of clarifying the debate, Cambridge Assessment, in conjunction with IPPR, explored the extent to which alternative assessment models can deliver on these important system objectives (Brooks and Tough 2006; Bell et al. 2008). This work resulted in three alternative

models for delivering these aims in the context of the English system. Whilst designed for England, these may also help with strategic and practical development of arrangements in other national systems.

Model 1: validity in monitoring plus accountability to school level

The first of these models uses a matrix sampling model to moderate teacher assessments. The sampling frame is dependent on the size and number of schools in the system, and presupposes the capacity to implement systems of supported teacher assessment of every child, moderated by a 'light sample' of children within each school. National examinations provide information for progression into the labour market and higher education.

The aim of this approach is to collect data using a national monitoring survey and to use this data for monitoring standards over time as well as for moderation of teacher assessment. This would enable school performance to be measured for accountability purposes and would involve a special kind of criterion referencing known as domain referencing. Question banks would be created based on the curriculum with each measure focusing on a defined domain. A sample of questions would be taken from the bank and divided into a large population of small testlets (smaller than the current national tests). These would then be randomly allocated to each candidate in a school. Every question is therefore attempted by thousands of candidates, giving robust summary statistics and summary statistics on a large sample of questions. This means that for a particular year we might know, for example, that on average candidates can obtain 50% of the marks in domain Y.

The following year we might find that they obtain 55% of the marks in that domain. This therefore measures the change and no judgement about relative year-on-year test difficulty is required. Neither is there a need for a complex statistical model for analysing the data, although modelling would be required to calculate the standard errors of the statistics reported. However, with the correct design they would be superfluous because they would be negligible. It would be possible to use a preliminary survey to link domains to existing levels and the issue of changing test items over time could be solved by chaining and making comparisons based on common test items between years. Although each testlet would be an unreliable measure, it would be possible to assign levels to marks using a statistical method once an overall analysis had been carried out. The average of the testlet scores would be a good measure of a school's performance, given that there are sufficient candidates in the school. The appropriate number of candidates would need to be investigated.

The survey data could also be used to moderate teacher assessment by asking the teacher to rank order the candidates and to assign a level to each of them. Teacher assessment levels would then be compared with testlet levels and the differences calculated. It would not be expected that the differences should be zero, but rather that the need for moderation should be determined by whether the differences cancel out or not. Work would need to be done to establish the levels of tolerance and the rules for applying this process would need to be agreed. The school could have the option of accepting the statistical moderation or going through a more formal moderation process.

There would be a number of potential advantages deriving from the use of this model. Validity would be increased, as there would be greater curriculum

coverage. The data would be more appropriate for the investigation of standards over time. The test development process would be less expensive as test items could be re-used through an item bank, including past test items from national curriculum tests. There would also be fewer problems with security related to 'whole tests'. No awarding meetings would be needed, as the outcomes would be automatic and not judgemental. Since candidates would not be able to prepare for a specific paper, current forms of negative wash-back and narrowing of the curriculum would be eliminated. There would also be less pressure on the individual student since the tests would be low stakes. Given that there are enough students in a school, the differences in question difficulty and pupil question interaction would average out to zero leaving only the mean of the pupil effects. From the data, it would be possible to generate a range of reports, e.g. equipercentiles and domain profiles. Reporting of domain profiles would address an issue raised by Tymms (2004) that 'the official results deal with whole areas of the curriculum but the data suggest that standards have changed differently in different sub-areas'.

Work would need to be done to overcome a number of potential disadvantages of the model. Transparency and perception would be important and stakeholders would need to be able to understand the model sufficiently to have confidence in the outcomes. This would be a particularly sensitive issue as students could be expected to take tests that prove to be too difficult or too easy for them. Some stratification of the tests according to difficulty and ability would alleviate this problem. There is an assumption that teachers can rank order students and this would need to be further explored. Applying the model to English would need further thought in order to accommodate the variations in task type and skills assessed that arise in that subject area. Eventually the model would offer the possibility of reducing the assessment burden but the burden would be comparatively greater for the primary phase. Although security problems could be alleviated by using item banking, the impact of item re-use would need to be considered. Having (active) test items in the public domain would be a novel situation for almost any other important test in the UK (except the driving test).

Discussion and research would be needed in a number of areas:

- values and validity;
- scale and scope, e.g. number and age of candidates, regularity and timing of tests;
- formal development of the statistics model;
- simulation of data (based on APU science data initially);
- stratification of tests/students;
- pilots and trials of any proposed system.

Model 2: validity in monitoring plus a switch to 'school-improvement inspection'

The second of the models relies on national school inspection arrangements to provide accountability of schools, with teacher assessment providing information for parents and children. National examinations provide information for progression into the labour market and higher education, while a light-sampling, low-stakes monitoring survey provides robust information on national standards.

Processes for equating standards over time in current arrangements pose significant challenges:

- teacher confidence in test outcomes;
- evidence of negative wash-back into learning approaches;
- over-interpretation of data at pupil group level; inferences of improvement or deterioration of performance not being robust due to small group size;
- ambiguity in policy regarding borderlining;
- publishing error figures for national tests.

In the face of these problems, it is attractive to adopt a low-stakes, matrix-based, light-sampling survey of schools and pupils in order to offer intelligence to UK government on underlying educational standards. With a matrix model underpinning the sampling frame, far wider coverage of the curriculum can be offered than with current national testing arrangements.

However, if used as a replacement for national assessment of every child at ages 7, 11 and 14, then key functions of the existing system would not be delivered:

- data reporting, to parents, progress for every child at the end of each key stage;
- school accountability measures.

In a system with a light-sampling model for monitoring national standards, the first of these functions could be delivered through (1) moderated teacher assessment, combined with (2) internal testing, or tests provided by external agencies and/or grouped schools arrangements. The DfES prototype work on assessment for learning (DCSF 2008) could potentially provide national guidelines for (1) the overall purpose and framework for school assessment, and (2) model processes. This framework of assessment policy would be central to the inspection framework used in school inspection.

The intention would be to give sensitive feedback to learners and parents, with the prime function of highlighting to parents how best to support their child's learning. Moderated teacher assessment has been proven to facilitate staff development and effective pedagogic practice. Arrangements could operate on a local or regional level, allowing transfer of practice from school to school.

The second of these functions could be delivered through a change in the Ofsted inspection model. A new framework would be required since the current framework is heavily dependent on national test data, with all the attendant problems of the error in the data and instability of standards over time. Inspection could operate through a new balance of regional/area inspection services and national inspection – inspection teams operating on a regional/area basis could be designated as 'school improvement teams'. To avoid competition between national and regional inspection, national inspections would be joint activities led by the national inspection service. These revised arrangements would lead to increased frequency of inspection (including short-notice inspection) for individual schools and increased emphasis on advice and support to schools in respect of development and curriculum innovation. Inspection would continue to focus on creating high expectations, meeting learner needs, and ensuring progression and development.

Model 3: adaptive, on-demand testing using IT-based tests

The third model relies on the development of a national infrastructure delivering electronic, on-demand, adaptive tests. This provides information back to teachers, pupils and parents. Data are built up in each school until a point is reached where there is a robust reflection of the performance of the school across the whole curriculum – this would be more frequently available from big schools and less frequently available from small schools. These constant data-feeds from schools would contribute to an ever-growing body of national data on underlying standards in the education system.

In 2002, Bennett outlined a new world of adaptive, on-demand tests, which could be delivered through machines. He suggested that 'the incorporation of technology into assessment is inevitable because, as technology becomes intertwined with what and how students learn, the means we use to document achievement must keep pace'. Bennett (2001) identifies a challenge, 'to figure out how to design and deliver embedded assessment that provides instructional support and that globally summarises learning accomplishment'. He is optimistic that 'as we move assessment closer to instruction, we should eventually be able to adapt to the interests of the learner and to the particular strengths and weaknesses evident at any particular juncture ... '. This is aligned to the commitments of UK government to encourage rates of progression based on individual attainment and pace of learning rather than age-related testing. In the government's five-year strategy for education and children's services (DfES 2004), principles for reform included 'personalisation and choice as well as flexibility and independence'. The White Paper on 14–19 Education and Skills (DfES 2005) stated, 'Our intention is to create an education system tailored to the needs of the individual pupil, in which young people are stretched to achieve, are more able to take qualifications as soon as they are ready, rather than at fixed times ... ' and 'to provide a tailored programme for each young person with intensive personal guidance and support'. These intentions are equally important in the context of national testing systems.

The process relies on item-banking, combining items in individual test sessions to feed to students a set of questions appropriate to their stage of learning and to their individual level of attainment. Frequent, possibly weekly, low-stakes assessments could allow coverage of the curriculum over a school year. Partial repetition in tests, whilst they are 'homing in' on an appropriate testing level, would be useful as a means of checking the extent to which pupils have really mastered and retained knowledge and understanding.

Pupils would be awarded a level at the end of each key stage based on performance on groups of questions to which a level has been assigned. More advantageously, levels could be awarded in the middle of the key stage as in the revised Welsh national assessment arrangements.

Since tests are individualised, adaptivity helps with security, with manageability and with reducing the 'stakes', moving away from large groups of students taking a test on a single occasion. Cloned test items further help security. This is where an item on a topic can include, for example, different number values on a set of variables, allowing the same basic question to be systematically changed on different test administrations, thus preventing memorisation of responses. A simple example of cloning is where a calculation using a ratio can use a 3:2 ratio in one test item version and a 5:3 ratio in another. The calibration of the bank would be crucial with

test item parameters carefully set and research to ensure that cloning does not lead to significant variations in test item difficulty.

Reporting on national standards for policy purposes could be delivered through periodic reporting of groups of cognate test items. As pupils nationally take the tests, and when a critical nationally representative sample on a test is reached, this would be lodged as the national report of standards in a given curriculum area. This would involve grouping key test items in the bank, e.g. on understanding a two-dimensional representation of three-dimensional objects, accumulating pupils' performance data on an annual basis (or more or less frequently, as deemed appropriate) and reporting on the basis of key elements of maths, English etc. This 'cognate grouping' approach would tend to reduce the stakes of national assessment, thus gauging more accurately underlying national standards of attainment. This would alleviate the problem identified by Tymms (2004) that 'the test data are used in a very high-stakes fashion and the pressure created makes it hard to interpret those data. Teaching test technique must surely have contributed to some of the rise, as must teaching to the test'.

Data could be linked to other cognate groupings, e.g. those who are good at X are also good at Y and poor on Z. Also, performance could be linked across subjects.

There are issues of reductivism in this model as there could be a danger to validity and to curriculum coverage, as a result of moving to test forms that are 'bankable', work on-screen and are machine-markable. It is certainly not the case that these testing technologies can only utilise the most simple multiple-choice (MC) test items. It must be noted that MC items are used as part of high-level professional assessment, e.g. in the medical and finance arenas, where well-designed test items can be used for assessing how learners integrate knowledge to solve complex problems (AAT 2009; Leighton and Gierl 2007).

However, it is certainly true that, at the current stage of development, this type of approach to delivering assessment cannot handle the full range of test items that are currently used in national testing and national qualifications. The limitation on the range of test item types means that this form of testing may best be used as a component in a national assessment model, and not the sole vehicle for all functions in the system.

School accountability could be delivered through this system using either (1) a school accumulation model, where the school automatically accumulates performance data from the adaptive tests in a school data record which is submitted automatically when the sample level reaches an appropriate level in each or all key subject areas, or (2) the school improvement model outlined in model 2 above.

There are significant problems of capacity and readiness in schools and it remains to be seen whether these can be swiftly overcome or are structural problems, e.g. schools adopting very different IT network solutions and arranging IT in inflexible ways (EU 2004). However, it is very important to note that current arrangements in England remain dominated by 'test sessions' of large groups of pupils, rather than true on-demand, adaptive tests. These latter arrangements could relieve greatly the pressures on infrastructure in schools, since sessions would be arranged for individuals or small groups on a 'when ready' basis.

There are technical issues of validity and comparability to be considered. The facility of a test is more than the sum of the facility on the individual test items that make up each test. However, this is an area of intense technical development in the assessment community, with new understanding and theorisations of assessment

emerging rapidly (Gustafsson 2005; Cohen et al. 2008). There are issues of pedagogy. Can schools and teachers actually manage a process where children progress at different rates based on on-demand testing? How do learners and teachers judge when a child is ready? Will the model lead to higher expectations for all students, or self-fulfilling patterns of poor performance amongst some student groups? These – and many more important questions – indicate that the assessment model should be tied to appropriate learning and management strategies, and is thus not neutral technology, independent of learning.

Exploring further alternatives ...

Alternative models that rely entirely on teachers assessing their own students against the levels of the national curriculum have their own difficulties. The major problem is that of reliability – a critical issue in any system that depends entirely on human judgement (Wood 1993; Klenowski 2006; Harlen 2004). It is important to recognise the reality – both technical limitations as well as putative benefits – of a national assessment system that depends entirely on teacher judgements. This would add significantly to the pressure on teachers (with key objections emerging from teacher unions in respect of the increased workload posed by national schemes of teacher assessment; NUT 2008) and there would most likely arise significant tensions relating to the forms of data that would be required by government, particularly between national data and accountability demands and the kind of teacher assessment that teachers would most likely desire as an optimum approach to supporting teaching and learning. Under the political circumstances obtaining nationally and internationally, it would be naïve to believe that any new model would be adopted by UK government if it did not include rigorous accountability data, which enables school performance to be measured. The political imperatives in England are so strong in this respect that it must be an essential part of any new national assessment model. One of the reasons why England has been, and still is, locked into the national testing system is that no alternative has been devised that will provide a balanced and technically viable model in respect of data at the level of the individual, data at the level of the school and data at the level of national performance.

The reviews by Daugherty (Wales) (2004) and Tomlinson (England) (2004) asserted a need to increase the role of teacher assessment in national systems. Neither review presented evidence that teacher assessment can operate in such a way as to deliver stable assessment outcomes in a context of high stakes accountability arrangements. Indeed Sweden offers evidence to the contrary, with acute 'grade inflation' accompanying the introduction of national accountability systems in a system relying heavily on teacher assessment (Wikström 2005). The principal example of teacher assessment advocated by policy makers etc. (Queensland) has not yet integrated accountability arrangements, nor has it generated comprehensive data on standard reliability measures etc. (CERI 2005; Klenowski 2006). Classification error is thus difficult to establish – a crucial problem. While the enhancement of learning remains an apparent benefit of such arrangements, the introduction of teacher assessment into a context over-determined by high-stakes accountability arrangements remains highly problematic. What is needed is well-designed research on the technical characteristics of teacher assessment under different system conditions. Without this, a drive towards teacher assessment could well be a leap of faith, in the dark. This carries worrying ethical implications.

As discussed, difficulties arise when national test data are used for school-level performance tables and for information on standards over time, since the instruments are not optimised for the multiplicity of functions for which they are being used. The problems detected in the Massey Report (Massey et al. 2003) in respect of maintaining test standards over time in some key stage assessments remain as a significant challenge to external national testing in England, and in the light of a synthesis of studies in this area, Tymms (2004) concluded that 'statutory test data must not be used to monitor standards over time'. The Statistics Commission in England (2005) commented that 'the primary purpose of the key stage tests at ages 7, 11 and 14 years, is to measure the progress of individual pupils against the National Curriculum, not to measure aggregate standards over time'. That the key purpose of measuring standards over time is not being adequately met by current arrangements presents a serious challenge to policy-makers' assumptions about the system.

The range of possible models outlined earlier illustrates the fact that there are a number of radically different possibilities that are worthy of consideration. Any new arrangements will require research and development with rigorous piloting and evaluation and the necessary ethical safeguards for learners involved in any development phase. Elsewhere, the authors have argued (Oates 2008) that the key issue of 'time' is crucial in understanding the limitations of policy formation, commenting that 'inadequate development time' and 'lack of adequate trialling/piloting' have been cited as factors contributing to severe defects and problems in a string of fundamental revisions to the education and training system. We go on to elaborate on how 'lack of time' compromises innovation and system improvement with little chance of building appropriate protocols for the enactment of effective public policy. To achieve a policy objective it is important to look at the different parts of the policy needs, e.g. to consider the legitimacy and manageability of the range of functions being attributed to a national assessment system, and to take seriously input from a range of stakeholders. If there is a commitment to develop robust systems, such development can take a long time, typically five years and upwards to develop an effective system. Of course, this is likely to be an unpalatable timescale for politicians who would prefer to have shorter timeframes in order to gain potential credit for their policy initiatives.

The gap between public understanding of assessment and expectations of technical rigour remains wide (Wood 1993; Newton 2005). Leading on from this, one key barrier to change is the scale of the shift needed to overhaul our current national testing arrangements, and the apparent simplicity of the current system itself, i.e. testing each and every child at the end of each key stage, with tests that cover a number of levels of attainment. Gaining public trust and confidence is crucial for any future large-scale developments and this may well be undermined if new arrangements were to be significantly more complicated than those currently in existence. However, a robust national assessment system that delivers on the three key levels of reporting that are required of it, will indeed most likely be more complex in form – generating significant issues in respect of professional and public understanding, and of professional and public trust.

The alternative models outlined in this paper are not exhaustive, nor are they intended to be. They are designed to be indicative of the existence of a range of alternatives – a wider range of forms than is currently being explored in the Making Good Progress pilots (SLTs) or other national initiatives such as Assessing Pupil Progress. The analysis here also suggests that piloting of more than one alternative

may well be prudent, and that full examination of fitness for purpose, wash-back effect and the full set of intended and unintended consequences will require carefully designed and extended trials. In the face of the scale of necessary development effort, and in the light of the growing impetus behind ethically based piloting (Oakley 1999, 2000; Oates 2008; AEA-E 2008), Paul Black's statement (Emeritus Professor of Science Education, King's College London – given at a Cambridge Assessment Research Seminar at the House of Commons, April 2007) provides both technical developers and policy-makers with a significant challenge: '... The basic premise underlying any good system is that it should do no harm and as much good as possible ...'.

References

AAT. 2009. http://www.aatglobal.com/AATmcg.htm (accessed 12 January 2009).

AEA-E. 2008. *Discussion Group 5. Ethics in assessment: Practice, research and guidelines*. Jannette Elwood and Tim Oates (United Kingdom). Association for Educational Assessment – Europe Annual Conference, 8 November 2008, Hissar, Bulgaria. http://www.aea-europe.net/userfiles/081014%20Programme.pdf

Arnott, M. 2004. *Reproducing gender*. New York: Routledge.

Assessment Reform Group. 2002. *Testing, motivation and learning*. Cambridge: School of Education, University of Cambridge.

Barber, M., and D. Graham. 1993. *Sense, nonsense and the National Curriculum*. London: Falmer Press.

BBC News online. 2008. Pilot progress tests made easier. http://news.bbc.co.uk/1/hi/education/7246871.stm (accessed 11 April 2008).

Bell, J., T. Bramley, S. Green, and T. Oates. 2008. *Alternatives to national testing at KS1, KS2 and KS3*. Cambridge: Cambridge Assessment.

Bennett, R. 2001. How the Internet will help large-scale assessment reinvent itself. *Education Policy Analysis Archives* 9, no. 5.

Bennett, R. 2002. Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning and Assessment* 1, no. 1.

Black, P., and D. Wiliam. 1998. Assessment and classroom learning. *Assessment in Education* 5, no. 1: 7–74.

Broadfoot, P., A. Pollard, M. Osborn, E. McNess, and S. Triggs. 1998. Categories, standards and instrumentalism: Theorizing the changing discourse of assessment policy in English primary education. Paper presented at the AERA conference, April, in San Diego.

Brooks, R., and S. Tough. 2006. *Assessment and testing: Making space for learning*. London: IPPR.

Cambridge Assessment. 2007. *Response to consultation: Making good progress*. Cambridge: Cambridge Assessment.

Cambridge Assessment MEI. Undated. Making good progress. MEI response to DfES consultation. http://www.mei.org.uk/files/pdf/MakingGdProgress_MEIResponse.pdf (accessed 12 January 2009).

Cambridge Assessment, National Foundation for Educational Research, and the Nuffield Foundation. 2009. *Policy and research seminar on National Assessment Arrangements for Key Stage 3, 9 January 2009, at The Nuffield Foundation – Proceedings*. Slough: National Foundation for Educational Research.

CERI. 2005. *Formative assessment: Improving learning in secondary classrooms*. Paris: Organisation for Economic Co-operation and Development.

Chitty, C. 1999. *The education system transformed*. Manchester: Baseline Book Company.

Clarke, C. 2003. Untitled speech delivered at National College for School Leadership conference, November 13. http://www.dcsf.gov.uk/speeches/media/documents/New%20Heads%20-%20SoS.20.11.03.doc (accessed 12 January 2009).

Cohen, J., T. Chan, T. Jiang, and M. Seburn. 2008. Consistent estimation of Rasch item parameters and their standard errors under complex sample designs. *Applied Psychological Measurement* 32, no. 4: 289–310.

Daugherty, R. 2004. *Learning pathways through statutory assessment: key stages 2 and 3. Interim report of the Daugherty Assessment Review Group*. Cardiff: Daugherty Assessment Review Group.
DCSF (Department for Children, Schools and Families). 2008. http://nationalstrategies.standards.dcsf.gov.uk/node/64401?uc=force_uj
Department for Education and Skills. 2004. *Five year strategy for children and learners*. Nottingham: DfES Publications.
Department for Education and Skills. 2005. *14–19 Education and skills: White Paper*. February. London: DfES.
Dobson, J., and K. Henthorne. 1999. Pupil mobility in schools. Research Brief Number 168, University College London.
EU. 2004. *Implementation of 'Education and training 2010' work programme – working group C 'ICT in education and training' progress report*. European Commission.
Gipps, C., and H. Goldstein. 1983. *Monitoring children: An evaluation of the assessment of performance unit*. London: Heinemann.
Gustafsson, J. 2005. The Rasch model in vertical equating of tests: A critique of Slinde and Linn. *Journal of Educational Measurement* 16, no. 3: 153–8. Published online: 12 September 2005.
Hansard. 2008. *National assessment – Maladministration*, 10 March 2008: Column 172W.
Harlen, W. 2004. Rethinking the teacher's role in assessment. Paper presented at the British Educational Research Assessment Annual Conference, September 16–18, University of Manchester.
Hopkins, D., and J. Sebba. 1995. Improving schools: an overview of improving the quality of education for all projects. Paper presented at the European Conference on Educational Research, University of Bath.
House of Commons Children, Schools and Families Committee. 2008. *Testing and assessment. Third report of session 2007–08, Volume I*. London: HMSO.
James, M. 2006. Assessment, teaching and theories of learning. In *Assessment and learning*, ed. J. Gardner. London: Sage Publications.
James, M., and C. Gipps. 1998. Broadening the basis of assessment to prevent the narrowing of learning. *Curriculum Journal* 9, no. 3: 285–97.
Klenowski, V. 2006. *Evaluation report of the pilot of the 2005 Queensland assessment task (QAT)*. Townsville, Queensland: James Cook University.
Leighton, J.P., and M.J. Gierl. 2007. *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge: Cambridge University Press.
Mansell, W. 2007. *Education by numbers – The tyranny of testing*. London: Politico's.
Massey, A. 2005. *Comparability of national tests over time: A project and its impact*. Research Matters, 1, 2–6.
Massey, A., S. Green, T. Dexter, and L. Hamnett. 2003. *Comparability of national tests over time: KS1, KS2 and KS3 standards between 1996 and 2001: Final report to QCA*. Cambridge: University of Cambridge Local Examinations Syndicate.
Monseur, C., H. Sibberns, and D. Hastedt. n.d. *Equating errors in international surveys in education*. University of Liège.
National Assessment Authority. 2008. http://www.naa.org.uk/single_level_tests/ (accessed 11 April 2008).
Newton, P. 2005. The public understanding of measurement inaccuracy. *British Educational Research Association Journal* 31: 419–42.
Newton, P. 2007. *Evaluating assessment systems*, Paper 1, June 2007. London: QCA.
Newton, P. 2008. *Monitoring national attainment standards – A collection of working papers*. Coventry: Office of the Qualifications and Examinations Regulator.
NFER. 2007. *Response to Making Good Progress consultation*. Slough: National Foundation for Educational Research.
NUT. 2008. NUT Briefing on the assessment for learning strategy and assessing pupils' progress initiative, September 2008. National Union of Teachers.
Oakley, A. 1999. Paradigm wars: Some thoughts on a personal and public trajectory. *International Journal of Social Research Methodology* 2, no. 3: 247–54.
Oakley, A. 2000. *Experiments in knowing*. Cambridge: Polity Press.

Oates, T. 2007. Bring back the APU! Presentation to Institute of Educational Assessors annual conference, May 3 2007, in London. http://www.ciea.org.uk/news_and_events/annual_conference/day1.aspx

Oates, T. 2008. Going round in circles: Temporal discontinuity as a gross impediment to effective innovation in education and training. *The Cambridge Journal of Education* 38, no. 1: 105–20.

Ofqual. 2008. *The regulation of examinations and qualifications: An international study*. London: Ofqual.

Ofsted. 2002. *The curriculum in successful primary schools*. Manchester: Office for Standards in Education.

Oxfam. 2008. The world we're in: Building a curriculum fit for the Twenty First Century. https://www.oxfam.org.uk/education/policy/england/files/DCSF%20NC%20Inquiry%20-%20Oxfam%20GB%20submission.pdf (accessed 21 January 2009).

Postlethwaite, N. n.d. *What do international assessment studies tell us about the quality of school systems?* University of Hamburg.

Statistics Commission. 2005. *Measuring standards in English primary schools,* Report no. 23, February. London: Statistics Commission.

Stobart, G. 2008. *Testing times: The uses and abuses of assessment*. Abingdon: Routledge.

Sutherland, S. 2008. *The Sutherland inquiry: An independent inquiry into the delivery of National Curriculum tests in 2008. A report to Ofqual and the Secretary of State for Children, Schools and* Families. London: HMSO.

Thomas, S., P. Sammons, P. Mortimore, and R. Smees. 1997. Stability and consistency in secondary schools' effects on students GCSE outcomes over three years. *School Effectiveness and School Improvement* 8: 169–97.

TGAT (Task Group on Assessment and Testing). 1988. *National Curriculum: A report*. London: DES.

Tomlinson, M. 2004. *14–19 Qualifications and curriculum reform*. Nottingham: DfES.

Tymms, P. 2004. Are standards rising in English primary schools? *British Educational Research Journal* 30, no. 4: 479–93. Education and Science and the Welsh Office. http://www.kcl.ac.uk/content/1/c6/01/54/36/TGATreport.pdf

Wellcome Trust. 2008. *Perspectives on education: Primary science*. London: Wellcome Trust.

Wikström, C. 2005. Grade stability in a criterion-referenced grading system: the Swedish example. *Assessment in Education: Principles, Policy and Practice* 12, no. 2: 125–44.

Wiliam, D. 2001. *Level best*. London: Association of Teachers and Lecturers.

Wiliam, D. 2003. National curriculum assessment: How to make it better. *Research Papers in Education* 18, no. 2: 129–36.

Wiliam, D. 2008. Six degrees of integration: an agenda for joined-up assessment. Paper presented at the Chartered Institute of Educational Assessors' National Conference, April 2008.

Wood, R. 1993. *Assessment and testing*. Cambridge: University of Cambridge Local Examinations Syndicate.

Wößmann, L., E. Lüdemann, G. Schütz, and M.R. Wes. 2007. *School Accountability, Autonomy, Choice, and the Level of Student Achievement: International Evidence from PISA 2003*. Education Working Paper No. 13. Paris: Organisation for Economic Co-operation and Development.