
# Alternative Approaches to National Assessment at KS1, KS2, and KS3

Sylvia Green
John F Bell
Tim Oates
Tom Bramley

2008

University of Cambridge
Local Examinations Syndicate

Cambridge Assessment

# Alternative Approaches to National Assessment at KS1, KS2, and KS3

## This paper

In 2006, Cambridge Assessment worked with the Institute for Public Policy Research (IPPR) to examine: (i) whether the original purposes of national assessment continue to be sensible aims; (ii) the extent to which the apparatus of national assessment arrangements in England are delivering on those objectives in a robust and cost-effective manner; and (iii) whether there are alternative models for national assessment arrangements which could be developed and trialled.

We scrutinised all extant evidence on current arrangements and identified the following set of benefits and deficits of the National Curriculum and its assessment regime:

### Benefits of the National Curriculum and its assessment

Entitlement (Chitty C; Colwill I)
Avoidance of repetition of content (Chitty C; Colwill I)
Progression (Chitty C; Sammons)
Balanced coverage in the primary phase (Sammons)
Help with pupil transfer (Dobson & Polley; Ewans)
Enhanced performance of girls (Elwood & Comber)
Enhanced development of skills (SCAA)
Identification of KS3 dip (INCA)
Higher expectations of young people (Hopkins, Barber)

### Deficits of the National Curriculum and its assessment

Acute overload (Alexander; Dearing)
Marginalisation of certain subjects (Rawling)
Overbearing assessment (Wiliam, Mansell)
Adverse impact of assessment on teaching and learning (Wiliam; Osborn; ARG)
Problems in maintaining standards over time (Massey; Tymms; Statistics Commission)
Failure to provide robust information on national standards (Oates, Tymms, Stats Commission)
Rise of instrumentalism in learners and 'maladministration' amongst teachers (ARG; QCA)

We observed that the benefits were principally associated with the provision of the National Curriculum, and that the deficits were principally associated with national test arrangements. As a result of our analysis we concluded, in respect of the assessment regime for the National Curriculum:

That the original purposes of national assessment are viable and should be retained, and can be restated succinctly as: *providing formative assessment for teaching and learning; providing information for school accountability; and providing information on national standards.*

That, whilst there have been benefits yielded by national assessment and attendance policy instruments such as performance tables, the weight of evidence now suggest that the benefits are outweighed by serious structural problems such as shallow learning styles amongst learners, test preparation dominating teaching; and a failure of the assessment system to provide the State with reliable evidence on national standards.

That there is a range of alternative models which should be developed and trialled. The trials should be designed using more robust approaches that those currently being used by DCSF for the development of the Single Level Tests, and should include effective ethical safeguards for all participants.

## Objectives

It seems extraordinary that, despite overwhelming evidence, there has been little consideration of the range of alternative models which might be adopted to better deliver the aims of national assessment. Over the last decade, obvious problems inherent in existing arrangements have been identified by a range of official reviews, measurement experts and researchers, but little effort has been devoted to implementing their recommendations. One reason for this may be that governance and accountability has been neglected by the QCA. Should it not be the case that the best minds in assessment should constantly review the technical performance of the national assessment system? Instead, some of the best assessment experts have been relegated to being contractors to the NAA in the development and administration of the tests – thus introducing a prima facie conflict of interest for them. Public scrutiny and public accountability mechanisms have thus been grossly deficient, allowing the accumulation of serious problems in national assessment arrangements.

The only alternative to existing arrangements which is being explored by the DCFS is the Single Level Test model – scrutiny of this model by both independent measurement experts and National Assessment Agency test developers has highlighted major weaknesses in the fundamental design principles of Single Level Tests. After a year of trialling, a stable test form with robust measurement characteristics still seems a long way off.

Single Level Tests should not be seen as the only alternative to existing arrangements – there are many alternatives which can start out with a far greater chance of meeting the three major objectives of national assessment. This paper is intended to explore just *some* of the possible options which are able to deliver on the key objectives of national assessment testing.

We believe that it is essential to develop and trial, in parallel, a variety of models which show promise, rather than develop only one alternative. We believe that such models should be subject to thorough critique, and any option which moves forward as a national approach should enjoy the support of all participants in the system.

In developing new arrangements, the aim should be to

- reduce the assessment burden on schools
- provide formative assessment for teaching and learning
- provide information for school accountability
- provide information on national standards.

It is vital to drive revised arrangements not only through concerns regarding the defects of existing arrangements, but also through a firm re-statement of educational purpose (values) and a commitment to high degrees of validity.

This paper does not outline values and validity in detail, but recognises that this is an essential precondition of designing revised arrangements, putting them in place, and monitoring their operation. It is vital that a full discussion of these matters precedes any executive decision regarding revised arrangements.

Here, we simply wish to state that any system should seek to:

- produce high quality information on trends in underlying attainment of pupils, in order to inform effective policy and action at national policy level
- encourage a relationship between assessment and learning whereby learning is enhanced through the operation of assessment
- support school improvement processes

This paper seeks to highlight different directions which could be taken in developing revised arrangements which both address known problems of existing arrangements and deliver on the three key objectives outlined immediately above.

A small note on sampling approaches for national standards (rather than simply adding up all the results of every child). A sampling method is far more robust and can cover all areas of the National Curriculum, unlike the limited coverage of each year's current national tests. But the call by many to 'bring back the Assessment of Performance Unit' (which sampled about 10% of children nationally at a time) is naïve, since the APU was not the infallible instrument which people now claim it to be. Lessons can be drawn from the troubled history of the APU and a new, far more robust model and processes developed, which – unlike current arrangements - would yield reliable evidence of trends in national standards.

## The characteristics of current national tests

Currently, a single test is administered, in each subject and at each key stage, to the national cohort of learners. National Curriculum Levels are determined by standard setting procedures. There are a number of problems with the current model.

In some cases the tests are not long enough to provide information to justify choosing cut-scores between adjacent marks even though the difference between adjacent marks can have a significant effect on the percentages of the cohort achieving particular levels. There are problems with misclassification of levels applied. Wiliam (2001a, 2001b) reports that 'it is likely that the proportion of students awarded a level higher or lower than they should be because of the unreliability of the tests is at least 30% at key stage 2 and may be as high as 40% at key stage 3'. Criterion referencing fails to work well since question difficulty is not solely determined by curriculum content. It can also be affected by 'process difficulty' and/or 'question or stimulus difficulty', (Pollitt et al,1985). It is also difficult to allocate curriculum to levels since questions testing the same content can cover a wide range of difficulty. There are other issues such as the impact of a narrowed curriculum based on what can be assessed in a paper and pencil test, the difficulty that parents have in understanding the levels and the timing of the tests at the end of a key stage - which means that the data cannot be used effectively. In the final report of the Assessment Review Group in Wales, Daugherty (2004) recommends that 'serious consideration should be given to changing the timing of Key Stage 3 statutory assessment so that it is completed no later than the middle of the second term of Year 9'. The Group believed the current timing to be unhelpful in relation to a process that could, in principle, inform,' and that, 'one source of information that would be of use potentially to pupils and their parents is not available until after the choice of pathway for Year 10 and beyond has been made'. There are also implications for the potential use of Key Stage 1 and 2 data for transition between phases. Difficulties also arise when national test data are used for league tables and for information on standards over time since the instruments are not designed to fulfil all of the functions for which they are being used. Massey et al. (2003) and Massey (2005) detected problems in maintaining test standards over time in some key stage assessments and Tymms (2004) concludes that 'statutory test data must not be used to monitor standards over time'. The Statistics Commission (2005) commented that 'the primary purpose of the key stage tests is to measure the progress of individual pupils against the National Curriculum, not to measure aggregate standards over time'. The Commission went on to conclude that, 'Key Stage test scores may not be an ideal measure of standards over time, but it does not follow that they are a completely unsuitable measure for a PSA target. There is no real alternative at present to using statutory tests for setting targets for aggregate standards'.

## Alternative models for national assessment

*Model 1*
*Validity in monitoring plus accountability to school level*

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 lots of small testlets (smaller than the current KS tests). These would then be randomly allocated to each candidate in a school. Every question is therefore attempted by thousands of candidates so the summary statistics are very accurate and there are summary statistics on a large sample of questions. This means that for a particular year we 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 items over time could be solved by chaining and making comparisons based on common 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 related to 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 items could be re-used through an item bank, including past 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 the 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 suggests 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 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 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'*

Whilst the processes for equating standards over time have been enhanced since the production of the Massey Report, there remain significant issues relating to:

- 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
- no provision to implement Massey recommendations regarding keeping tests stable for 5 years and then 'recalibrating' national standards
- 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 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 testing of every child at the end of KS1, 2 and 3, 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 (i) moderated teacher assessment, combined with (ii) internal testing, or tests provided by external agencies and/or grouped schools arrangements. The DfES prototype work on assessment for learning could form national guidelines for (i) the overall purpose and framework for school assessment, and (ii) 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*

In 2002, Bennett outlined a new world of adaptive, on-demand tests which could be delivered through machines. He suggests 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 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 (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 and 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 items further help security. This is where an item on a topic can include 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 ratio can use a 3:2 ratio in one item version and 5:3 ratio in another. The calibration of the bank would be crucial with item parameters carefully set and research to ensure that cloning does not lead to significant variations in item difficulty.

Reporting on national standards for policy purposes could be delivered through periodic reporting of groups of cognate 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 area. This would involve grouping key items in the bank e.g. on understanding 2D representation of 3D objects and 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 that 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, eg 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 curriculum coverage as a result of moving to test forms which are 'bankable', work on-screen and are machine-markable. Using the Cambridge taxonomy of assessment items is one means of monitoring intended and unintended drift. It is certainly not the case that these testing technologies can only utilise the most simple multiple-choice (mc) items. MC items are used as part of high-level professional assessment e.g. in the medical and finance arenas, where well-designed items can be used for assessing how learners integrate knowledge to solve complex problems.

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 items which are currently used in national testing and national qualifications. The limitation on the range of item types means that this form of testing is best 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 (i) 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 (ii) the school improvement model outlined in model 2 above.

There are significant problems of capacity and readiness in schools, as evidenced through the problems being encountered by the KS3 ICT test project which has successively failed to meet take-up targets. 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. However, it is very important to note that current arrangements remain based on 'test sessions' of large groups of pupils, rather than true on-demand, adaptive tests. These 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 items which 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.

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.

## Discussion

Each of the models addresses the difficulties of multipurpose testing. However, each model also presents challenges to be considered and overcome. The Statistics Commission (2005) commented that 'there is no real alternative at present to using statutory tests for setting targets for aggregate standards'. The task is to find such an alternative. The real challenge is to provide school accountability data without contaminating the process of gathering data on national standards and individual student performance. All three models have significant advantages over both existing national test arrangements and the proposed Single Level Tests and could lead to increased validity, reliability and utility in national assessment arrangements.