How to promote educational quality through national assessment systems

Sylvia Green and Tim Oates
Cambridge Assessment

A paper presented at the International Association for Educational Assessment Annual Conference, Baku, Azerbaijan, September 2007.

Sylvia Green
Director, Research Division
Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU
UK
Direct dial. +44 (0)1223 553844
Fax. +44 (0)1223 552700
Email:
green.s@cambridgeassessment.org.uk
www.cambridgeassessment.org.uk

Tim Oates,
Group Director
Assessment Research & Development
Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU
UK
Direct dial. +44 (0)1223 552750
Fax. +44 (0)1223 552700
Email:
oates.t@cambridgeassessment.org.uk
www.cambridgeassessment.org.uk

How to promote educational quality through national assessment systems

Introduction

In our paper we address some of the challenges posed by the development of national assessment systems. We 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. We explore a range of key elements including,

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

We outline the dangers of multi-purpose testing in the context of political and educational objectives. 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. We conclude 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 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, with some amendments, has been with us since the early nineties. In 1988 a Task Group on Assessment and Testing was set up by the government of the day and chaired by Professor Paul Black of King's College, University of London. At the moment there is a sense that change could be possible or at least that there is a chance for real debate about the way forward. A number of organisations are calling for change, including the General Teaching Council, the national Union of Teachers and the Association of Teachers and Lecturers. If that is the case we must work to make the current system better and to make sure that we learn valuable lessons from the past. It is vital that the relationship between assessment and the impacts on learning should be fully understood so that any new model of national assessment can be designed to enhance that relationship.

It is essential to understand the pattern of incentives and drivers a specific model exerts on the system, in order to ensure that these pressures accord with public policy objectives.

The Task Group on Assessment and Testing was set up to advise on assessment and testing within the new national curriculum that had been introduced in England. The group's report placed a very strong emphasis on the formative purpose of national tests.

One of the proposals 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 the issue of publication 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 <u>only</u> 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 with summative functions incorporated 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.

One of the aims 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. The size and detail of the then new national curriculum created problems for the operationalisation of the TGAT objectives. There were initially 976 statements of attainment against which each eleven year old had to be assessed. Tick boxes reigned supreme and assessment went into a state of overload with an enormous increase in teachers' workload. The logistics of carrying out the practical activities and assessing each child against so many criteria led to a great deal of anxiety and stress amongst teachers and children.

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. National coherence improved, expectations were increased through a challenging curriculum and there were more opportunities for professional development of assessment expertise.

In terms of the 'standards agenda' a key question is – did the changes and initiatives introduced at that time lead to the gains in test performance that have been so widely celebrated? One of the major problems in attempting to answer that question is the difficulty of measuring the comparability of national test standards over time. A project was commissioned by the Qualifications and Curriculum Authority in England to investigate the comparability of test standards and was carried out by researchers at Cambridge Assessment. We investigated the stability of test standards at ages 7, 11 and 14 years of age in English, maths and science from 1996 to 2001 and reported varied findings across age groups and subjects with some tests appearing more lenient and some more severe over time. The extent of the variation immediately raised questions in respect of the claims being made for significant gains in national attainment. It is clear that there were gains but not to the extent indicated by the top-line figures for the national tests from 1996 to 2001.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 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. Negative effects occur when too much time is spent on memorisation, question spotting and test practice to the detriment of positive teaching and learning. 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' (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.

A key problem with our current system has been identified and commented upon in many contexts and on many occasions and that is 'fitness for purpose'. The debate on this issue has led to an accumulation of evidence and a review of this issue is long overdue. There has been a continual process of incremental modification which has mainly involved 'bolting on' additional requirements, for example, in relation to optional tests and additional national data collection. 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. This makes it a huge task to effect real change to the system since we are 'locked into' it in so many ways. We need to separate the functions and recognise that the current instruments used to assess the individual student and inform learning are not adequate for national monitoring nor for calling 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 because the accountability function impedes the ability to use assessment as an integral part of the learning process, thus placing the teacher in a context with strongly competing imperatives.

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. 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.

However, such studies are not unproblematic as instruments for national governments interested in close and accurate scrutiny of the performance of their own systems. Although wide-ranging, these studies are not undertaken at optimum times for national systems, for example, nations may be introducing innovations which require system-level monitoring. Other problems are that: such studies are insufficiently frequent for national monitoring purposes; they are prone to 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 curriculum and assessment arrangements in specific national settings.

Nation-specific processes of system monitoring thus 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, and the Assessment of Performance Unit (APU) – now redundant, National Curriculum Assessment and Performance Tables – all in England, monitoring in the Scottish system, in New Zealand and in new systems being introduced in Germany and Australia.

NAEP and APU are similar 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 time frames), 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 school 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.

The National Curriculum and Performance Table models in England rely on using data from each and every child (from national tests taken at 7, 11 and 14, and national subject-based examinations taken at 16 and 18) to build a national picture, to provide information for learners, parents and teachers and to judge the performance of schools. The APU model, described above, was abandoned when national testing was introduced into England by law in 1988. The consensus amongst researchers is that the current national testing system in England has serious structural defects, mainly relating to the problem of using data from national tests and examinations for too many conflicting purposes. Major problems are attached to: 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; misclassification of attainment in terms of 'levels'; and acute effects of 'teaching to the test' rather than focusing on 'deep learning'.

As an educational organisation, one of Cambridge Assessment's aims is to influence thinking on assessment, both nationally and internationally, and within our research division we have expertise in test development, national assessment models and national monitoring.

We have recently reviewed work on national assessment and national monitoring arrangements and have devised three alternative models for securing three essential aims.

Our expression of the fundamental aims of national assessment arrangements are 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

With the Institute for Public Policy Research (a UK Government think tank), Cambridge Assessment has developed three alternative models for delivering these aims, in the context of the English system. Whilst designed for England, these systems may have application in other national systems.

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

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 as well as offering a valid picture of the attainment standards and pattern of improvement (or decline) in individual schools.

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 and the issues that arise in any system that depends entirely on human judgement. It is important to recognise the reality of a national assessment system that depended entirely on teacher judgements. This would add significantly to the pressure on teachers and there would inevitably be problems related to the kind of data that would be required by government with potential tension between the national data and accountability demands and the kind of teacher assessment that teachers would want to carry out for teaching and learning purposes. It would be naïve to believe that any new model would be adopted if it did not include rigorous accountability data so that school performance can 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 we have been, and still are, locked into the national testing system is that no alternative has been devised that will provide data at the level of the individual, of the school and at the level of national performance.

As discussed, difficulties arise when national test data are used for league tables and for information on standards over time, since the instruments are not capable of fulfilling 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) 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'.

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. Oates highlights the key issue of 'time', 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' (2008, *forthcoming*). He goes 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; for example, to consider the range of functions of a national assessment system, and to take seriously input from a range of stakeholders. Such development can take a long time, as long as five years 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.

Another 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. assessing every child at ages 7, 11 and 14. Gaining public trust and confidence is crucial for any future large scale developments and this might be undermined if new arrangements were to be more complicated than those in existence. However, a robust national assessment system that delivers on the three key levels of reporting that are required of it, will be complex, since there are complex technical measurement issues involved. In this context we invoke the sentiment of

H.L. Mencken, ‘For every complex problem there is an answer that is clear, simple, and wrong.’

Any effective national assessment system should be designed to be fit for purpose and supportive of teaching and learning – that should be the overall goal. After almost twenty years’ experience we must have accumulated evidence about national assessment. The fundamental lesson that we must incorporate into any new thinking is that the use of national test data for multiple purposes has a negative impact. We must find ways of dealing with each function effectively and without negative wash-back. At Cambridge Assessment we have investigated a number of options and from the work we’ve done so far we are convinced that the challenge is not insurmountable. Any new, enhanced system should not only build on what has gone before but should also overcome the problems that exist in our current system.

Cambridge Assessment remains highly engaged not only with the operation of assessment systems, but also with the uses of assessment data for vital national purposes. This work on performance measurement indicates the importance that Cambridge Assessment attaches to the enhancement of learning arrangements and the management of national systems. The models outlined are not exhaustive; we welcome discussion of arrangements which contribute to the enhancement of systems in varying national contexts.

And the final word from Professor Paul Black, who commented at a recent Cambridge Assessment seminar in the House of Commons in London,

> *The basic premise underlying any good system is that it should do no harm and as much good as possible.*

References

Assessment Reform Group (2002). *Testing, Motivation and Learning.* Cambridge: School of Education, University of Cambridge.

Massey, A., Green, S., Dexter, T. and Hamnett, L. (2003). *Comparability of national tests over time: KS1, KS2 and KS3 standards between 1996 and 2001.* Final report to QCA. Cambridge: UCLES.

Massey, A. (2005). Comparability of national tests over time: A project and its impact. *Research Matters: A Cambridge Assessment Publication*, *1*, 2-6.

Oates, T. (2008). In: *Knowledge, Transformation and Impact*, Special Issue. *The Cambridge Journal of Education,* March 2008 *(forthcoming)*

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

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

Tymms, P. (2004). Are standards rising in English primary schools? *British Educational Research Journal*, *30*, 4, 479-493.