Cambridge Assessment

National Foundation for Educational Research

The Nuffield Foundation

Policy and research seminar on

National Assessment Arrangements for Key Stage 3

Friday 9 January 2009, at The Nuffield Foundation

Proceedings

Edited by Marian Sainsbury and Sarah Maughan

Contents

Introduction

Towards the end of 2008, national assessment policy in England became subject to rapid change. On 14 October, the Secretary of State for Children, Schools and Families announced the immediate end to national tests at Key Stage 3, a system that had been in place since the mid-1990s. In the same speech, an Expert Group was announced with the purpose of proposing revised arrangements, reporting in February 2009.

This seminar for assessment specialists took place on 9 January, 2009 and was designed to assemble as much advice and intelligence as possible to support the development of new arrangements.

The basis of discussion was a belief that the recently-announced changes to KS3 assessment arrangements confronted many of the structural problems which had accumulated in the system. However, the suddenness of the announcement left a question as to what arrangements should replace them, and when. The Expert Group would face considerable challenge in terms of the urgency of their work and the seminar was designed to feed into their advisory processes. It tackled:

- the underlying principles and purposes which should drive the revised arrangements;
- the range of alternative models which might exist and should be explored;
- the shortcomings of previous attempts to put in place a sample-based methodology for monitoring national standards, including a look at international experiences;
- issues around teacher assessment;
- the nature of the development process for new arrangements.

Contributions were made by Chris Whetton (National Foundation for Educational Research), Tim Oates (Cambridge Assessment) and Gordon Stobart (Institute of Education, representing the Nuffield-Foundation-sponsored Assessment Reform Group).

This report summarises the presentations and discussions. The seminar operated under the Chatham House rule, but the three presenters have given permission for their views to be attributed.

Presentation 1
International comparisons and the APU: lessons learned

Chris Whetton, NFER

This presentation appears in full as Appendix 2.

Summary of presentation

A national monitoring survey is one possibility being explored for enabling system-wide accountability after the removal of the end of Key Stage 3 tests. In this presentation, Chris Whetton addressed in some depth the nature of actual national and international monitoring surveys, highlighting the different approaches and the advantages and problems associated with each. The broad method for such surveys is to assess a sample of pupils drawn to be representative of the population and thereby to monitor standards of attainment within that population. In a matrix sampling design, not all pupils take the same tests, allowing wide coverage of the curriculum.

The Assessment of Performance Unit (APU) ran in England from 1974 to 1989. The original purpose was to identify underachievement, but this was never adequately defined, so the main outcomes were explorations of aspects of the curriculum. The tests were practical as well as written, addressing pupils aged 11, 13 and 15 years. There was some disagreement about statistical issues at the time, as well as significant problems of administration, conflict between developers and policy makers, and escalating sample size (see the NFER case study on APU also submitted to the Expert Group).

In the United States, there has been a **National Assessment of Educational Progress (NAEP)** since 1969. The ages sampled are nine, 13 and 17 years and the assessments take the form of written tests. The stated purposes of the NAEP assessment have changed and become broader in the course of its life. Its role has grown so that it now monitors achievement in individual states as well as for the nation.

The **Scottish Survey of Achievement (SSA)** has run since 2005, replacing the previous Assessment of Achievement Programme which started in 1983. It includes practical and investigative tests as well as written. Particular features include the intention to establish a relationship with teacher assessment, and the involvement of teachers throughout the process to develop assessment expertise in the workforce.

In New Zealand, the **National Education Monitoring Project (NEMP)** has an explicit aim of supporting teachers in using assessment for formative purposes. There is a strong emphasis on practical tasks and group work, using video for analysis. There is no accountability purpose in this system and it does not really perform the function of measuring performance over time.

International surveys include **TIMSS** (mathematics and science) and **PIRLS** (reading), assessments organised by the IEA, and **PISA**, run by the OECD, which aim to measure readiness for adult life. In some countries, these surveys are used as the only national monitoring system, sometimes with expanded samples.

The **purposes** for which national monitoring surveys are used emerged across this presentation as a frequent cause of difficulty, and NFER's final recommendations include the clarification of purposes in advance of setting up any new survey.

Once purposes are clear, then due attention must be given to the nature of the **sample** and the **tests**, to the **analysis** approach and to **reporting**. These generally interact together and decisions about one influence the others. NFER's full recommendations appear as Appendix 3.

Presentation 2
Where things are: a personal view

Tim Oates, Cambridge Assessment

This presentation appears in full as Appendix 4.

Summary of presentation

Tim Oates started by reflecting upon the benefits and deficits of the national curriculum with its associated assessment arrangements, highlighting the continuing need for a review of the curriculum as well as its assessment. It is clear that there remain problems in respect of the coherence of the national curriculum (Oates references Schmidt in his presentation) and its continuing bulk. Dylan Wiliam's emphasis on deep conceptual understanding and on the failure of teachers to be able to construct appropriate questions to probe this conceptual understanding are also critical issues. Oates went on to discuss the functions of a national assessment system.

The importance of agreeing the purpose(s) for a new system up front was emphasised. At the outset of the national curriculum the TGAT report identified four functions of assessment: formative, summative, evaluative and informative. Of these, the evaluative, or accountability, function soon took prominence but the informative purpose was also important in establishing an understanding of the new national curriculum. Currently, the system has many purposes – for example, accountability of schools and departments, upwards pressure on standards and tracking of individual progress. These purposes are intertwined, as are the agencies, such as local authorities and Ofsted, that make use of the assessment outcomes.

In response to this confusion, Cambridge Assessment proposes three basic purposes. National assessment arrangements should provide:

1. formative assessment for teaching and learning;
2. information for school accountability;

3. information on national standards.

To achieve these, three possible models have been identified:

Model 1: **Monitoring plus accountability to school level.** Teacher assessments collected for the whole population would be moderated by a national sample. The national data would also be used to monitor standards over time.

Model 2**: Monitoring plus a switch to 'school improvement inspection'.** In addition to a sample survey to monitor national standards, Ofsted would be reconstituted so that accountability at the school level would be provided through the inspection system. Information for parents and children would be provided by teacher assessment.

Model 3: **Adaptive, on-demand testing using IT-based tests.** This system would provide accountability through the accumulation of data. The data would be available to provide information to pupils, parents and teachers. It would also, over time, be accumulated to provide information at the school level and the national level. These three models were developed as practical options and to demonstrate that many models are possible in addition to the current 'candidates' of single level tests and Assessing Pupil Progress.

If Cambridge Assessment and IPPR can come up with three models which meet the three aims regarding feedback to parents, teachers and pupils, accountability and monitoring national standards, then there surely is a large range of possible models. There should be parallel trialling of a variety of models which enables us to understand the unintended consequences of any given model. This would avoid Oates' paradox in developing new models for national assessment: '…don't make the mistake of comparing the known disadvantages of the old with the putative advantages of the new…'.

A research process incorporating parallel developments is required, together with robust ethical safeguards. Full support from all parts of the system should be secured when the new system is implemented. Any new model should, at the point of national launch, be the preferred option of all key stakeholders; this is not a context in which a model should be imposed.

Presentation 3
Teacher assessment and assessment for learning

Gordon Stobart, Institute of Education

This presentation appears in full as Appendix 5.

Summary of presentation

This presentation had two strands, with more emphasis on the role of teacher assessment in a renewed KS3 system than on assessment for learning, which is less central to the current debate.

The Assessment Reform Group definition of summative teacher assessment is:

The process by which teachers gather evidence in a planned and systematic way in order to draw out inferences about their students' learning, based on their professional judgment, and to report at a particular time on their students' achievements (ARG, 2006)

A crucial issue concerns the **purposes** of assessment, making the current KS3 debate particularly interesting. Two key determining purposes – school-level targets and evaluation of teacher performance – have been removed at a stroke.

As a consequence, the emphasis shifts to asking: what is possible for teacher assessment in a system where accountability has been removed? The new arrangements should avoid replacing national testing for accountability with a local version of the same thing.

In a system reflecting the new requirements, four possible approaches to KS3 assessment can be identified. Each would serve different purposes.

1. English, mathematics and science could be assessed by teachers in the same way as all other subjects. This would involve no particular requirement for moderation.
2. There could be a model with tests/item banks to validate teacher judgements, as at KS1, in Scotland or in New Zealand's ASTTLE system.
3. Teacher assessment might be moderated or quality assured for local accountability and transfer arrangements, as happens in Wales.
4. As the Diploma becomes established, there could be a need for a KS3 certificate, assessed by teachers against the common national curriculum, to precede transfer to multiple pathways post-14. As in point 3 above a system of moderation would be needed to support this.

Gordon Stobart concluded by briefly discussing **assessment for learning**. This should be seen as day-to-day classroom assessment and should be established as part of good teaching practice. To complement this, there is a role for a summative teacher assessment system suited to a 'big ideas' curriculum at KS3.

Discussion summary

The discussion was lively and wide-ranging, covering questions of broad principle as well as practical issues for the short and medium term. The main points raised are brought together here by theme, rather than in the order in which they were raised; where there was a general consensus, this is indicated.

General Aspects of the New System

Purposes of assessment

The fundamental question of the purposes of assessment, which emerged as a theme from all three of the introductory presentations, also figured largely in the discussions. There was a consensus on the importance of establishing the purpose and function of the new KS3 arrangements in advance of setting up systems.

The three-purpose model proposed in Tim Oates's presentation drew widespread support. There was some doubt about how far the accountability purpose would, in fact, disappear from the new KS3 arrangements. Some tension was apparent between proposals that might be philosophically desirable and what might be politically and practically possible. The terms of reference for the Expert Group might suggest that decisions had already been made about restricting assessment purposes to providing information to others.

There was agreement that assessment purposes could not be seen in isolation, but impacted one on another. In light of this, it would be important to review the effects of decisions about one purpose on the others and to avoid unintended consequences.

There was general agreement about the need to be guided by consideration of the pupil's experiences during KS3 and for the overall effects of the assessment system to be, at the very least, benign towards learners. There was some feeling that improving the quality of education should be regarded as the principal purpose of the assessment system, but this did not command universal support, with others seeing assessment as the 'speedometer', rather than the driver, of educational progress.

There is a need to keep in mind the nature and purpose of KS3 itself and to reflect the approaches of the new KS3 curriculum and the characteristics of pupil progress over those three years. The positive practical task is to consider what good assessment should look like in this context. There is a need to help teachers to do a better job. In Wales, there has been much consideration of the purposes of KS3 and what the most pressing problems actually are. The emerging view is that there were too many pupils not making enough progress, and that the end of KS3 marked the end of a broad common curriculum, and therefore, end-of-KS3 assessment should document learners' achievements at this point (a view that reflected one of the suggestions made earlier by Gordon Stobart).

To have a positive impact upon pupil learning, whole-school measures would be inappropriate, as the focus must be upon the individual in the classroom.

Some commentators in the wider world believed that assessment should allow pupils to be compared one with another, or each with a standard, as an additional purpose.

Accountability

Within the wider area of assessment purposes, some discussion focused particularly upon the notion of school-level accountability.

In Northern Ireland, there is a strict separation between information collected for formative purposes and school accountability, with an agreement not to collate and aggregate data from formative profiles.

In Wales, the removal of the accountability function has led to a slight decrease in levels reported by teacher assessment, apparently showing that where teachers are trusted to exercise their own judgement, they place importance upon rigour and accuracy. Only in a high-stakes system did teachers feel that they were obliged to demonstrate high attainment.

In England, there were examples of anxiety leading to the adoption of accountability measures over and above what was required by policy. Some uses of Fischer Family Trust data are now having a negative backwash effect. Ten per cent of heads of department were reported to have pay rises dependent upon test results. Existing tests are being used even when not required.

In the United States, there have been reports that NAEP narrows the curriculum even though it is supposedly low-stakes.

The providers of tests and of data analysis should restrict the purposes for which they allow their outcomes to be used, and should specify that reports must include mention of measurement error. This could avoid the invalid use of inferences for accountability purposes. It would also be good practice to indicate the degree of uncertainty inherent in any result reported.

There is a public distrust of assessment results, related to statistics showing that standards have improved whilst employers' observations do not bear this out. At KS3, high stakes testing is not necessary as testing at KS4 fulfilled the accountability function. Teacher performance could be better judged by observation and in-school monitoring than by test results, provided the reliability of this process is assured. It is important to get the 'unit of interest' right. School accountability is not endorsed by contemporary research on school improvement. It is the interaction at classroom level, with the individual teachers, which should be the vital focus of attention.

Diversity

Several contributions questioned whether a universal compulsory system was the best approach as a replacement for KS3 tests, valuing instead a diversity of strategies. The Expert Group might establish broad principles, but allow different ways of fulfilling these.

The argument behind this was that since nobody could predict the outcomes and impacts of any new system introduced, it would be better to allow a number of different pilots as the system developed. Those systems that worked well could continue, with a variety in use at any time, whilst those that did not work would be discontinued at the end of a pilot period. Some government policies are built upon a valuing of diversity, with involvement in different ways by local organisations.

An assessment curriculum might exist at national or at school level, and in Finland local development , with direct involvement of teachers, has apparently been successful. Similarly, in the USA there is more tolerance of allowing different models to be suggested and evaluated.

Teachers

A number of contributions addressed the effects of assessment systems on teachers. It was felt that the teacher is the most important aspect of a pupil's education, rather than school level accountability, and there is a large amount of teacher variability. Any measure put in place to monitor teaching should be sensitive to the parts of the system that a teacher can have an impact on.

Teachers' anxieties were the result of their being evaluated, and of the profound mistrust of teacher judgement that characterised current policy. In a high-stakes system, teachers feel a need to meet demanding targets, putting them on the defensive. This could work against an approach aimed to benefit pupils. Teachers have been deprofessionalised by the current system, with its reliance on tests and lack of support for their own judgement, leading to a lack of confidence. However, the ability to assess is central to the teacher's role. The aftershocks on teachers of the removal of the KS3 tests have been interesting to observe with a mix of celebration and concern regarding target-setting and monitoring.

It was felt that one solution to the changes at KS3 would be to invest the budget in continuing professional development for teachers, to enable them to assess more reliably.

New technologies

The potential of new technologies to revolutionise assessment systems was highlighted.

There is a need to reflect the characteristics of 21st-century learning at KS3. Now a greater volume and quality of data could be preserved by schools in multimedia electronic formats. This would support stronger links between formative and summative assessment, and could be used with a matrix sampling design for policy audit. This new ability to sample behaviour that had been preserved electronically, allowing 'snapshots' of attainment, could remove

the need for 'psychometrics and flaky data'. This 'classroom aggregation technology' should be explored for a future totally different from present assumptions.

A further use of new technology could involve computer adaptive tests to deliver questions with a range of different difficulties to pupils, gradually settling on a level dependent on the questions that were answered correctly or incorrectly as the pupil proceeds through the test.

Research and evaluation

There was general agreement, sometimes strongly expressed, that high quality research and evaluation must be intrinsic to any new system. This must include full trialling, for demonstration of benefits and identification of unintended consequences. Innovation should be accompanied by a proper conceptual consideration of methods for achieving policy aims and evaluation of pilots before policy decisions are made.

Further, the research should be carried out using robust methodologies: the current tendency to use management consultants or market research organisations was considered to be highly deficient. It was emphasised that any new system must be properly planned, purposes agreed, pilots run, washback effects monitored, and so on, and that all of this takes time and should not be rushed.

Components of a New System

As described above it was suggested that the new system may have different purposes, at the different levels set out in Tim Oates' presentation. Revised models containing different elements of assessment need to be designed to meet these different purposes. The purposes must not be contradictory, and great care needs to be taken in aligning instruments and approaches to the purposes, whilst also ensuring that any instrument does not unintentionally impact on another purpose.

The discussion focused largely on the use of summative teacher assessment to provide information for pupils, parents and teachers, with assessment for learning as an integrated aspect of good teaching, and a national monitoring survey to monitor the system as a whole. The following sections describe the discussions that took place regarding the different aspects of this approach.

Summative Teacher Assessment

It was suggested that a system of summative teacher assessment could be introduced to replace a number of the functions of the KS3 tests, particularly in the area of providing information to pupils, parents and teachers. There was discussion of how to establish reliable teacher judgement as the main assessment approach.

Professional development was generally agreed to be crucial in establishing teachers' assessment skills. This might be supported by the introduction of Chartered Assessor status, school accreditation or licensing arrangements for assessors or moderators.

Systems where teachers learned from each other, with expert intervention, were commended as a focus for future development and funding. There was disagreement, however, about whether research yet showed that this approach had raised standards. A further view was expressed that there is no such thing as 'the characteristics of teacher assessment'. There are different characteristics under different system conditions, as each have different drivers, different incentives, different assessment models, etc.

The Assessing Pupils' Progress (APP) system, being piloted in England as the preferred method for establishing teacher assessment, drew some criticism. It has been described as assessment for learning, but lacks the genuine characteristics of this. Proper demonstrations of the validity and reliability of its outcomes are needed. There were some concerns about teacher workload.

One proposal, for England as in Wales, was to introduce a teacher-assessed KS3 certificate, to precede post-14 diversification.

Different subjects were considered to need different approaches, raising a question of how much should be required centrally. It was suggested that banks of optional tests or tasks could be made available to teachers to support them in making judgements. In Sweden, there are systems of moderating teacher assessments with externally set examination results, where teachers adapt elements of the national examination to align with the local curriculum. Teachers were invited to reconsider their assessments when they differed from the results.

In the US, research suggests that teachers and students 'conspire' together against rigour and quality by focusing on surface learning to optimise test scores, indicating that some external check is needed.

Do nothing

An alternative to the summative teacher assessment approach was to do nothing to replace KS3. As there is system accountability in secondary schools at the GCSE level, and teachers regularly assess pupils at the end of each academic year, a more formal national system may not be required. An alternative view was that a decision about this area should be postponed until the ongoing curriculum review is complete.

Other participants viewed the 'do nothing' suggestion as politically unrealistic, as the Government would be concerned about progress at KS3 and would want to replace the current system with something better. The remit of the Expert Group appeared to rule out 'doing nothing' as a possibility.

However, it was suggested that Governments should not feel they always have to have a right answer, and that they can ask the expert community to come up with answers. There was some support for this view.

Assessment for learning

These discussions addressed formative day-to-day assessment, involving learners in monitoring their own progress. It was agreed that this should be taking place irrespective of other decisions about assessment at KS3 and is largely independent of national monitoring or summative teacher assessment. This is just an ongoing aspect of good teaching and learning.

There was evidence that the best teachers were able to do this, fostering metacognition in their pupils, but general agreement that teachers need support in this area. Assessment information needs to be fine-grained in order to guide learning, and could be moment-to-moment, not just day-to-day. Research shows that teachers do not have well-enough developed theories of what is happening in pupils' heads, nor of focused intervention based on marking pieces of work. Support in assessment for learning might take the form of:

- developing teacher practices;
- probes designed for teacher use;
- or 'sharp tools' to focus assessment.

Feedback was crucial for assessment for learning, but expert/novice comparisons showed very different responses to the same feedback – 'throwing bottles in the sea' without knowing whether the feedback would be received or acted upon.

Much misunderstanding of assessment for learning still exists, and it is important that it should not be reified and institutionalised, but rather seen as a philosophical pedagogy pervading all work.

There was evidence that summative drives out formative, making it difficult to attempt to introduce a system that incorporates both: summative teacher assessment and assessment for learning systems ought to be kept separate.

Sample surveys

The question of whether to introduce a national monitoring survey did not appear to be controversial among this group of assessment specialists. Attention seemed to focus rather on the details of a possible system. A good deal of the discussion of sample surveys was concerned with the role of international surveys in contributing to national monitoring information.

One view was that participation in the international surveys made it unnecessary to introduce a national monitoring survey at all. International surveys found favour because they lead to a greater international awareness amongst teachers, raising questions about teaching and learning, whilst not directly influencing national policy. Other advantages are consistency and the approach to confidentiality.

Several points, both practical and theoretical, were adduced against the use of international surveys alone for monitoring purposes. Only one mathematics and science survey, TIMSS, directly addresses the KS3 age group, so nothing would be available for English. A further

practical point is the lengthy time delay between survey and report in international studies. However, this might not be a problem in terms of the monitoring purpose alone; it would become a problem if international results were required to be used for more immediate purposes, but the purposes should be kept distinct.

Educationally, there was a major problem of alignment between the content and processes assessed in international surveys and the national curriculum. International surveys were generally narrower and made no attempt to assess practical or exploratory work. It was suggested that if we were to place more emphasis on the results then greater involvement in the development process would be required to ensure some control over the content that is included in the tests. The new national curriculum focus on process skills should not be lost. National curriculum assessment had some advantages in helping teachers to understand the curriculum and the best of this should be preserved in any new system. Indeed, it could be said that it was the assessment system that enforced and embedded the national curriculum.

However, there was general agreement that there should be active consideration of integrating international surveys as part of national monitoring, keeping in mind the purposes of assessment, perhaps using some elements of the international surveys as part of a new national survey. This would allow links to be made from national standards as determined by the national survey, and international standards as determined by TIMSS.

If this course were to be adopted, there should be comprehensive and sophisticated involvement by the UK in the management of the international studies.

Concluding remarks

To round off the seminar, concluding remarks were offered by Tim Oates.

We are still currently in the midst of the review process, so no final conclusions can be drawn from the debate today. However, many valuable points have been raised and some of these can be highlighted.

The debate on purposes of assessment is fundamental, and it was clear from the discussions that this is not simply a matter of listing possible assessment purposes, but of a fundamental consideration of the nature and purpose of learning at KS3 in the context of the national curriculum, to ensure that the assessment system engages positively with this. It would be tempting, but wrong, to adopt a route that simply gives a national curriculum level in the least damaging way possible.

The arguments for doing nothing are quite persuasive. It is important to understand the complexity of the totality of the arrangements; getting the assessment right is not just about adopting a sound assessment model. This is essential but not sufficient. The totality of arrangements should work in concert in relation to purpose: assessment, national curriculum content, teacher training, accountability, etc. This makes it essential to

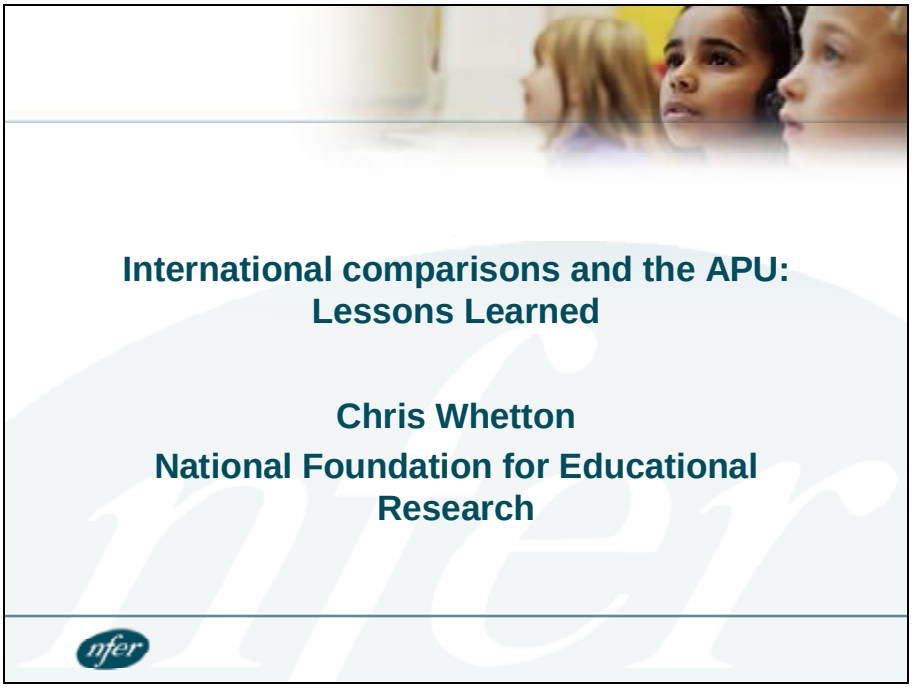
investigate unintended consequences; making it important to evaluate pilots before a full roll-out of new arrangements.

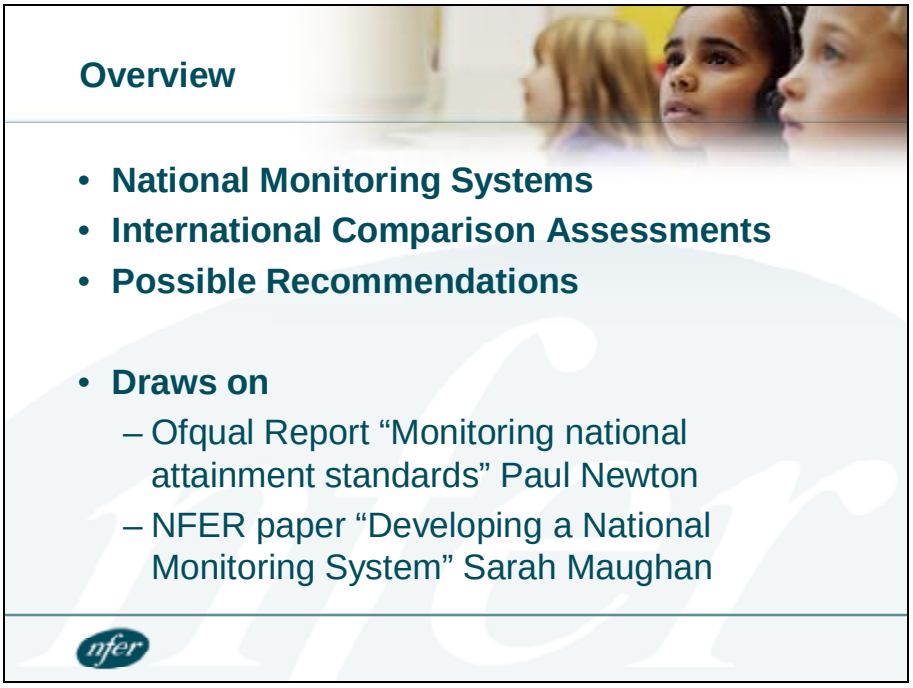
This seminar was set up to stimulate a debate about what is possible in the wake of the KS3 tests. The Proceedings from this seminar will be sent to the Expert Group.

Appendix 1: List of participants

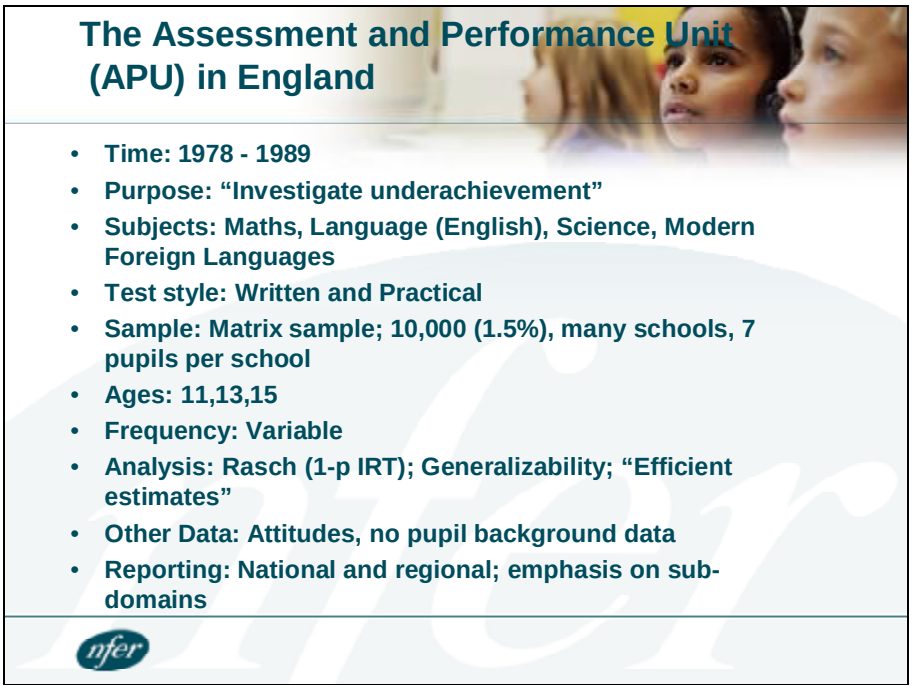
Dr Jo-Anne Baird	University of Bristol
Mr John Clay (on behalf of Ms Kathy Baker)	General Teaching Council
Mr Mike Baker	Education Journalist & Broadcaster
Mr John Bangs	National Union of Teachers
Ms Lorna Bertrand	Department for Children Schools and Families
Dr Mary Bousted	Association of Teachers and Lecturers
Professor Tim Brighouse	Institute of Education
Ms Audrey Brown	Department for Children Schools and Families
Professor Richard Daugherty	Cardiff University
Mr John Fairhurst	Association of School and College Leaders
Professor John Gardner	Queen's University Belfast
Mr Josh Hillman	The Nuffield Foundation
Professor Mary James	University of Cambridge
Mr Simon Lebus	Cambridge Assessment
Mr Warwick Mansell	Times Educational Supplement
Mr Simon Masterson	Department for Children Schools and Families
Ms Sarah Maughan	National Foundation for Educational Research
Ms Lydia Mulholland	Department for Children Schools and Families
Mr Tim Oates	Cambridge Assessment
Mr Dennis Opposs	Office of the Qualifications and Examinations Regulator
Ms Lesley Ravenscroft	Acumina
Professor Terry Russell	Centre for Research in Primary Science and Technology
Dr Marian Sainsbury	National Foundation for Educational Research
Professor Gordon Stobart	Institute of Education
Professor Peter Tymms	Durham University
Mr Chris Whetton	National Foundation for Educational Research
Professor Dylan Wiliam	Institute of Education

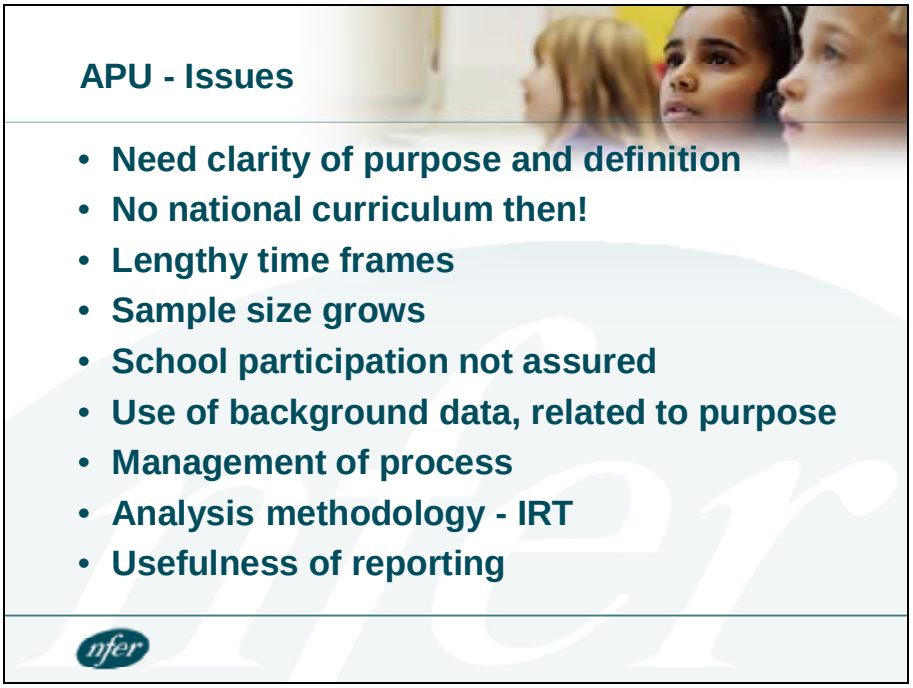
Appendix 2: Presentation by Chris Whetton

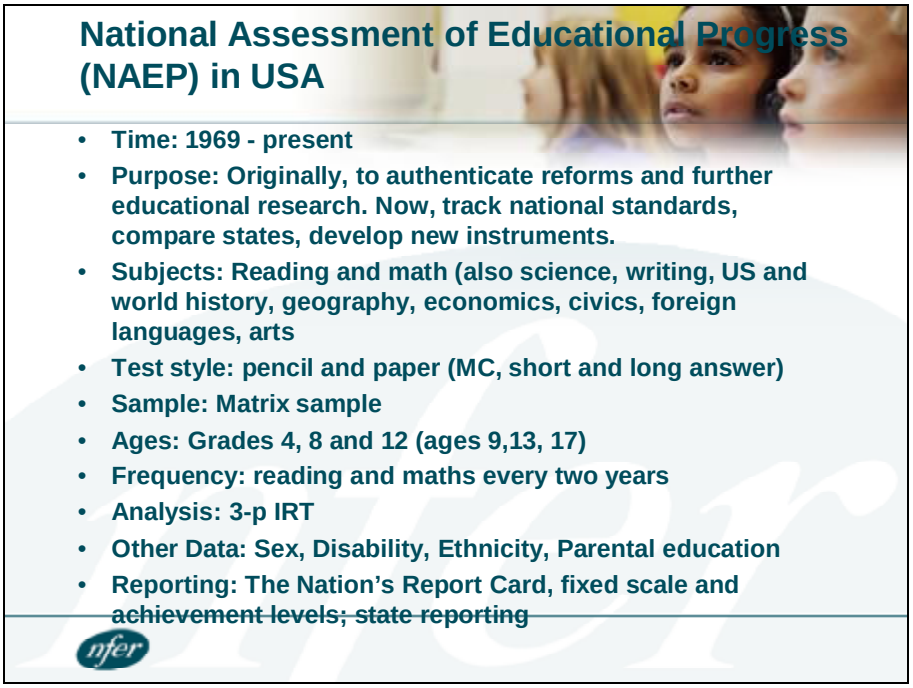


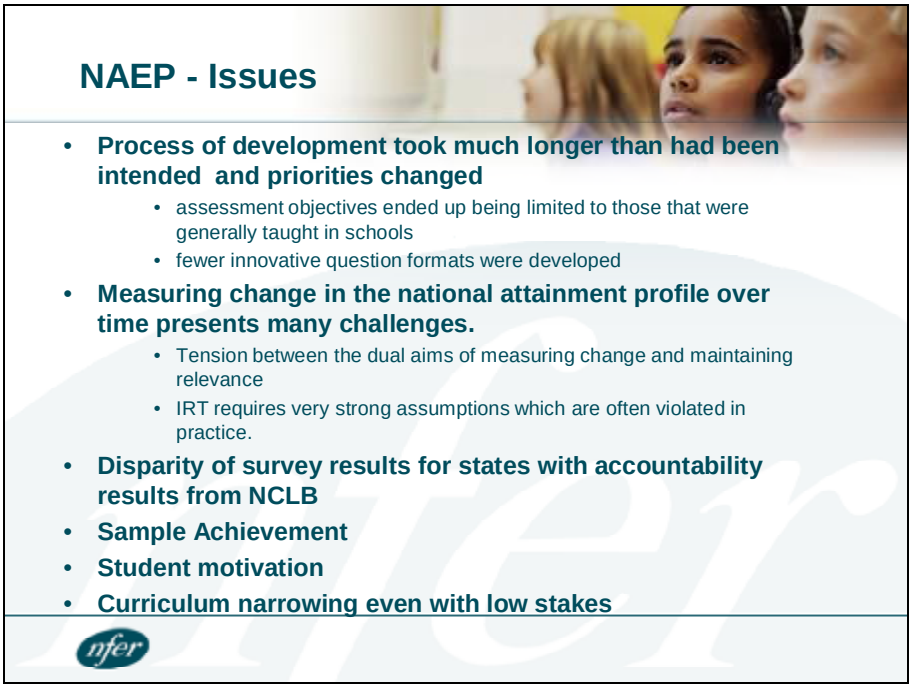


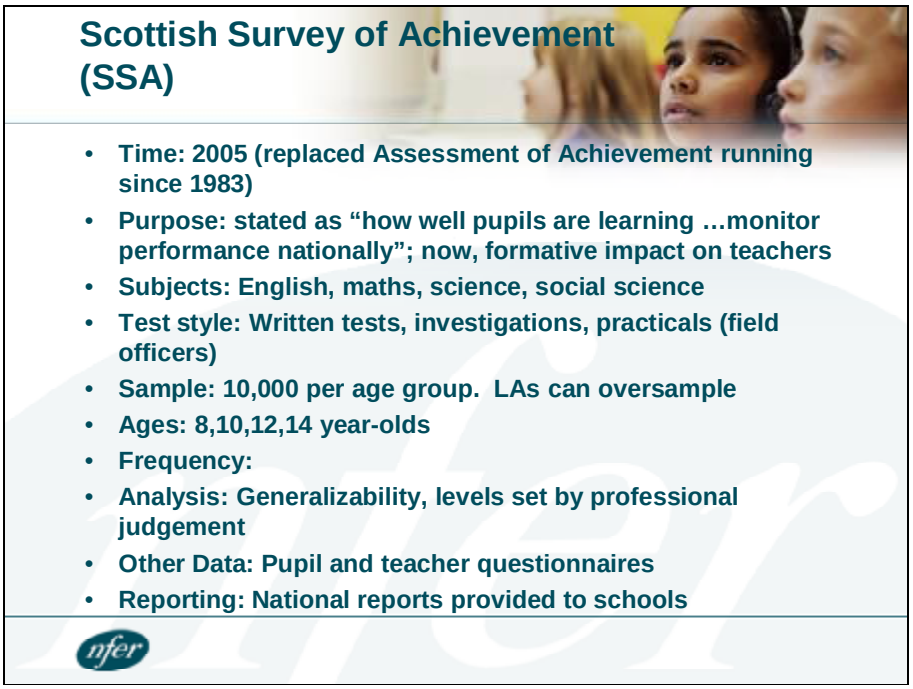

National Monitoring Systems
• The Assessment and Performance Unit (APU) in England
• National Assessment of Educational Progress (NAEP) in USA
• Scottish Survey of Achievement (SSA)
• National Education Monitoring Project (NEMP) New Zealand

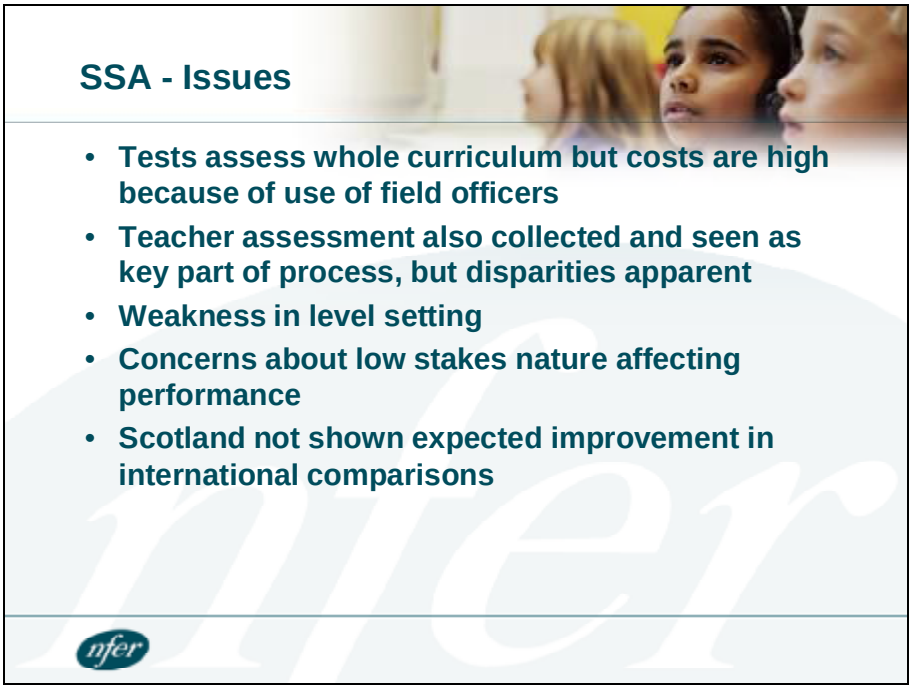

The Assessment and Performance Unit (APU) in England
• Time: 1978 - 1989
• Purpose: "Investigate underachievement"
• Subjects: Maths, Language (English), Science, Modern Foreign Languages
• Test style: Written and Practical
• Sample: Matrix sample; 10,000 (1.5%), many schools, 7 pupils per school
• Ages: 11,13,15
• Frequency: Variable
• Analysis: Rasch (1-p IRT); Generalizability; "Efficient estimates"
• Other Data: Attitudes, no pupil background data
• Reporting: National and regional; emphasis on sub-domains

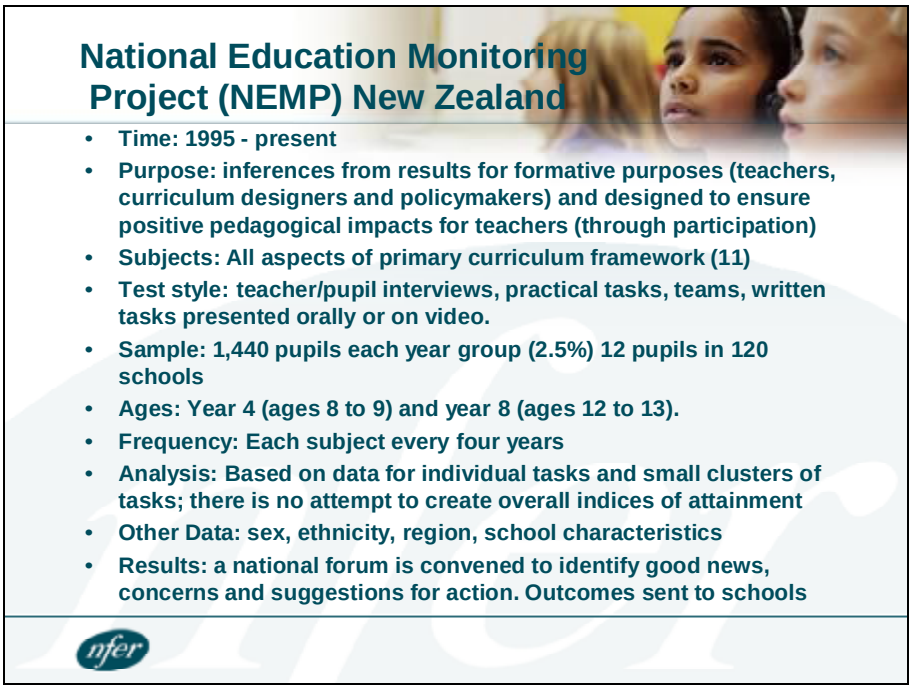

APU - Issues
• Need clarity of purpose and definition
• No national curriculum then!
• Lengthy time frames
• Sample size grows
• School participation not assured
• Use of background data, related to purpose
• Management of process
• Analysis methodology - IRT
• Usefulness of reporting
nfer

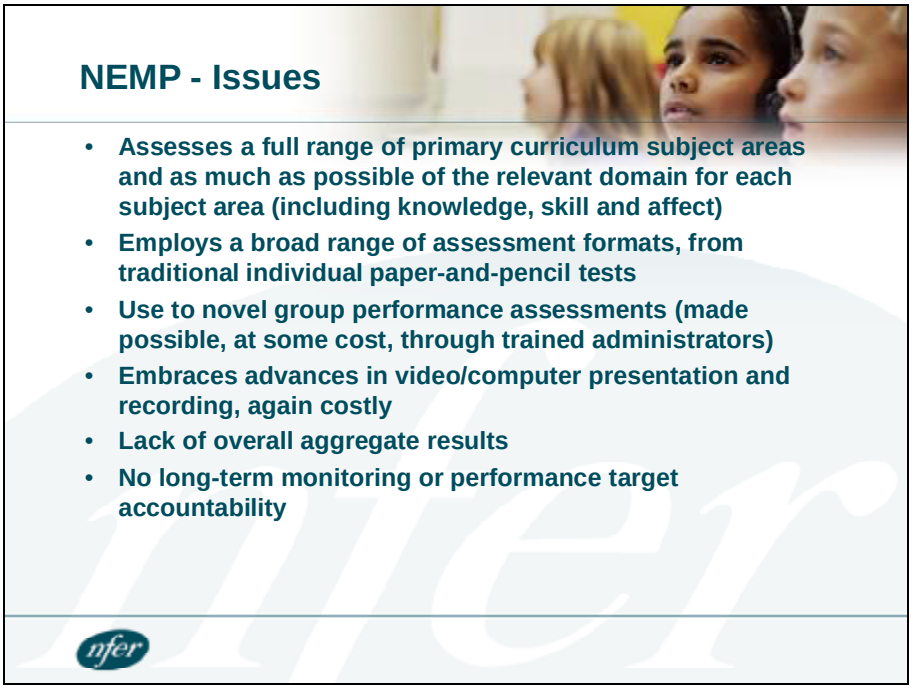

National Assessment of Educational Progress (NAEP) in USA
• Time: 1969 - present
• Purpose: Originally, to authenticate reforms and further educational research. Now, track national standards, compare states, develop new instruments.
• Subjects: Reading and math (also science, writing, US and world history, geography, economics, civics, foreign languages, arts
• Test style: pencil and paper (MC, short and long answer)
• Sample: Matrix sample
• Ages: Grades 4, 8 and 12 (ages 9,13, 17)
• Frequency: reading and maths every two years
• Analysis: 3-p IRT
• Other Data: Sex, Disability, Ethnicity, Parental education
• Reporting: The Nation's Report Card, fixed scale and achievement levels; state reporting
nfer

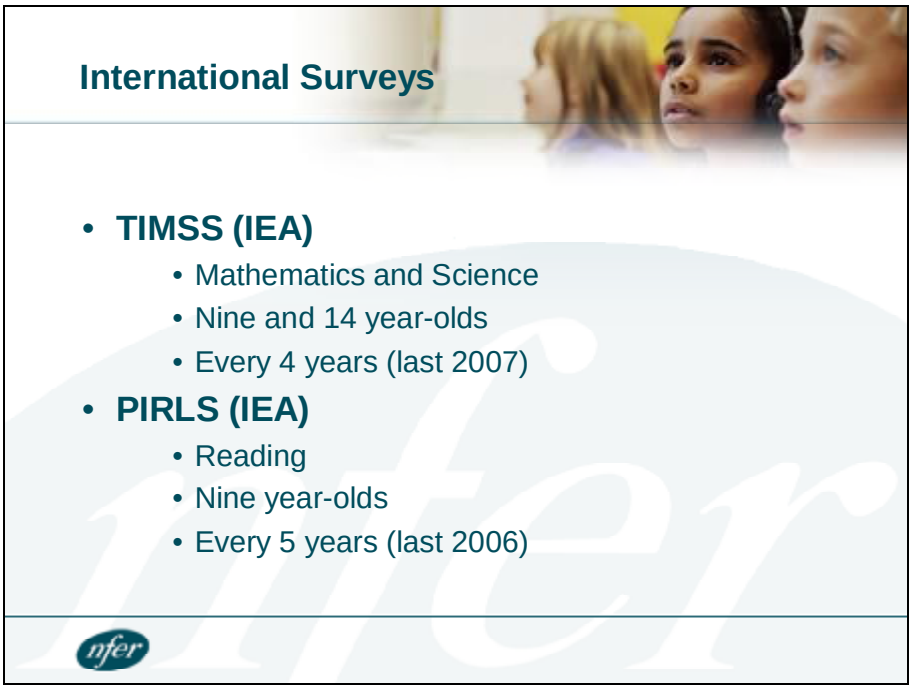

NAEP - Issues
• Process of development took much longer than had been intended and priorities changed
• assessment objectives ended up being limited to those that were generally taught in schools
• fewer innovative question formats were developed
• Measuring change in the national attainment profile over time presents many challenges.
• Tension between the dual aims of measuring change and maintaining relevance
• IRT requires very strong assumptions which are often violated in practice.
• Disparity of survey results for states with accountability results from NCLB
• Sample Achievement
• Student motivation
• Curriculum narrowing even with low stakes
nfer

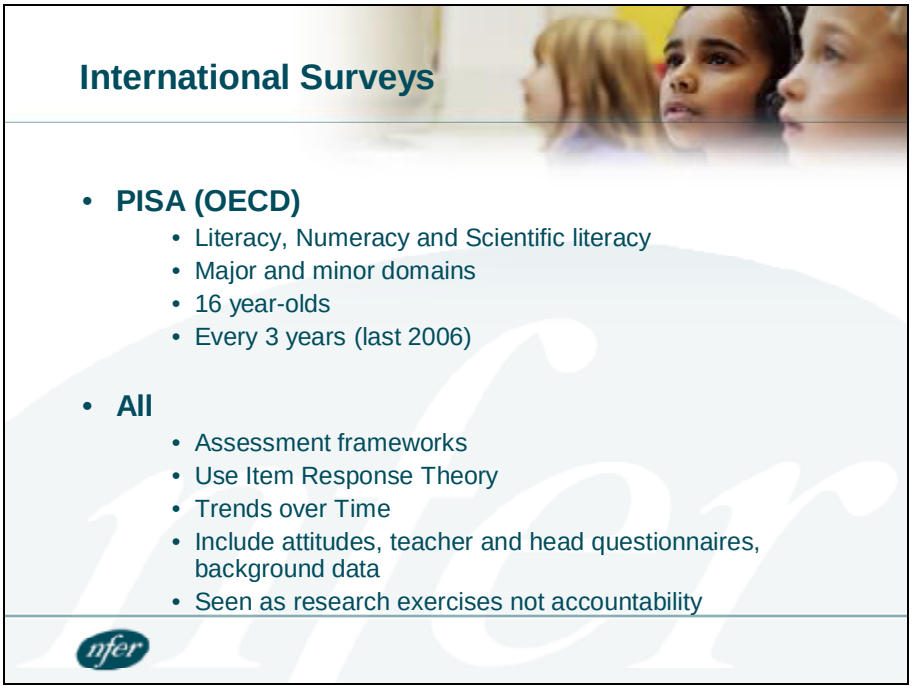

Scottish Survey of Achievement (SSA)
• Time: 2005 (replaced Assessment of Achievement running since 1983)
• Purpose: stated as "how well pupils are learning …monitor performance nationally"; now, formative impact on teachers
• Subjects: English, maths, science, social science
• Test style: Written tests, investigations, practicals (field officers)
• Sample: 10,000 per age group. LAs can oversample
• Ages: 8,10,12,14 year-olds
• Frequency:
• Analysis: Generalizability, levels set by professional judgement
• Other Data: Pupil and teacher questionnaires
• Reporting: National reports provided to schools
nfer

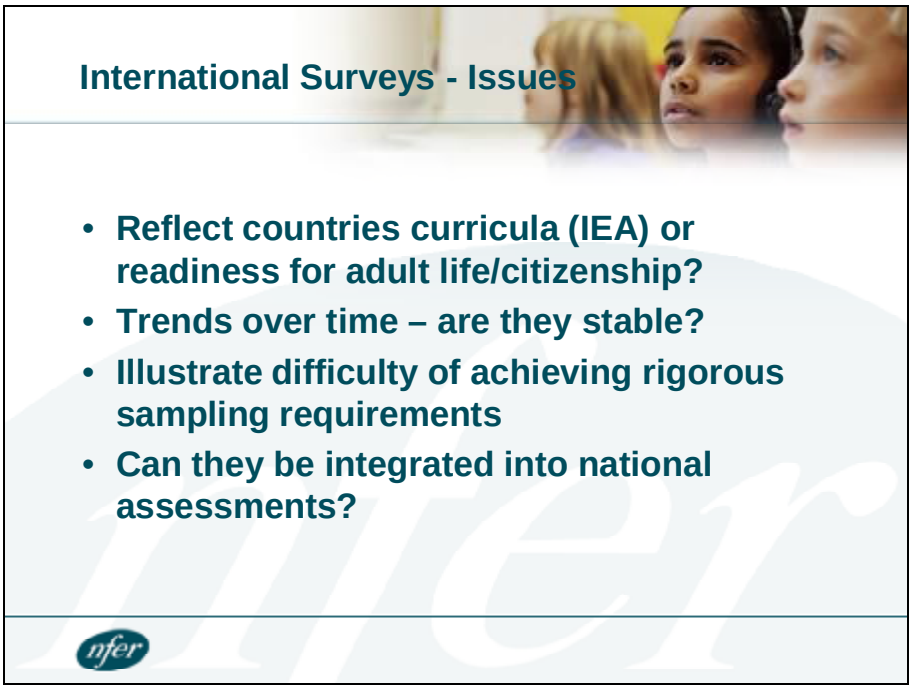

SSA - Issues
• Tests assess whole curriculum but costs are high because of use of field officers
• Teacher assessment also collected and seen as key part of process, but disparities apparent
• Weakness in level setting
• Concerns about low stakes nature affecting performance
• Scotland not shown expected improvement in international comparisons

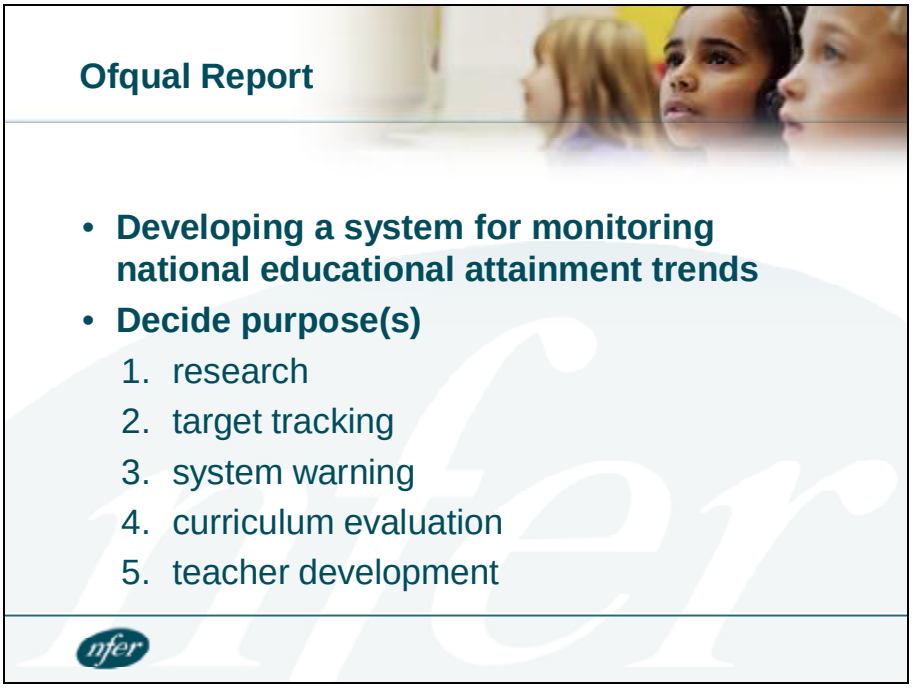

National Education Monitoring Project (NEMP) New Zealand
• Time: 1995 - present
• Purpose: inferences from results for formative purposes (teachers, curriculum designers and policymakers) and designed to ensure positive pedagogical impacts for teachers (through participation)
• Subjects: All aspects of primary curriculum framework (11)
• Test style: teacher/pupil interviews, practical tasks, teams, written tasks presented orally or on video.
• Sample: 1,440 pupils each year group (2.5%) 12 pupils in 120 schools
• Ages: Year 4 (ages 8 to 9) and year 8 (ages 12 to 13).
• Frequency: Each subject every four years
• Analysis: Based on data for individual tasks and small clusters of tasks; there is no attempt to create overall indices of attainment
• Other Data: sex, ethnicity, region, school characteristics
• Results: a national forum is convened to identify good news, concerns and suggestions for action. Outcomes sent to schools

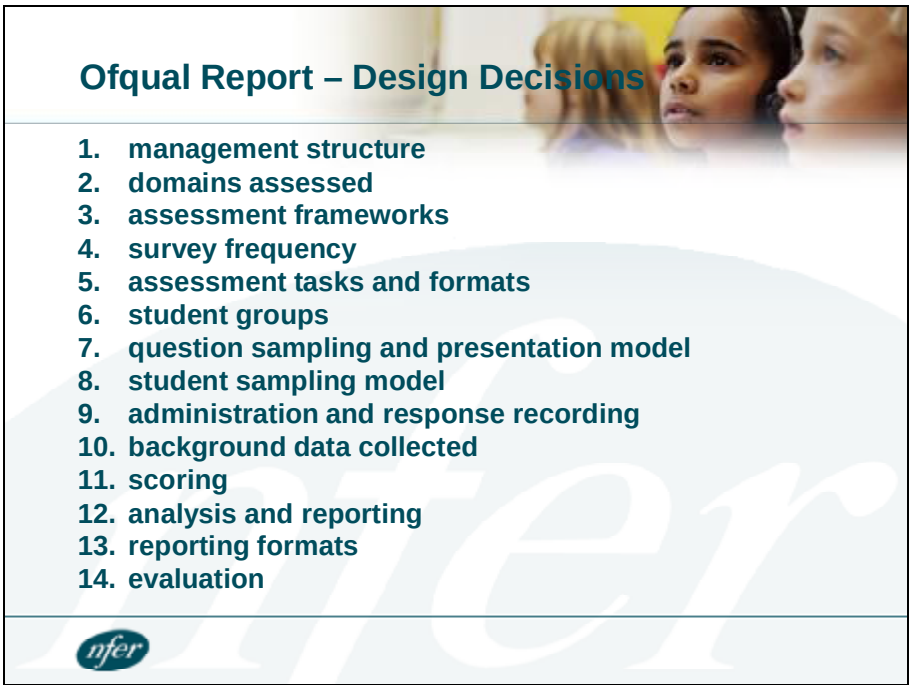

NEMP - Issues
• Assesses a full range of primary curriculum subject areas and as much as possible of the relevant domain for each subject area (including knowledge, skill and affect)
• Employs a broad range of assessment formats, from traditional individual paper-and-pencil tests
• Use to novel group performance assessments (made possible, at some cost, through trained administrators)
• Embraces advances in video/computer presentation and recording, again costly
• Lack of overall aggregate results
• No long-term monitoring or performance target accountability


International Surveys
• TIMSS (IEA)
• Mathematics and Science
• Nine and 14 year-olds
• Every 4 years (last 2007)
• PIRLS (IEA)
• Reading
• Nine year-olds
• Every 5 years (last 2006)


International Surveys
PISA (OECD)
Literacy, Numeracy and Scientific literacy
Major and minor domains
16 year-olds
Every 3 years (last 2006)
All
Assessment frameworks
Use Item Response Theory
Trends over Time
Include attitudes, teacher and head questionnaires, background data
Seen as research exercises not accountability
nfer


International Surveys - Issues
Reflect countries curricula (IEA) or readiness for adult life/citizenship?
Trends over time – are they stable?
Illustrate difficulty of achieving rigorous sampling requirements
Can they be integrated into national assessments?
nfer


Ofqual Report
• Developing a system for monitoring national educational attainment trends
• Decide purpose(s)
1. research
2. target tracking
3. system warning
4. curriculum evaluation
5. teacher development


Ofqual Report – Design Decisions
1. management structure
2. domains assessed
3. assessment frameworks
4. survey frequency
5. assessment tasks and formats
6. student groups
7. question sampling and presentation model
8. student sampling model
9. administration and response recording
10. background data collected
11. scoring
12. analysis and reporting
13. reporting formats
14. evaluation


NFER Recommendations
• Purposes
• The Sample
• The Tests
• Analysis
• Reporting


Department for Research in Assessment and Measurement
National Foundation for Educational Research
www.nfer.ac.uk
c.whetton@nfer.ac.uk

Appendix 3: NFER Recommendations

The way in which a national monitoring survey is set up will depend in large part on a number of key decisions that will need to be made in the early phases. These decisions will impact on all future decisions, and as such the following recommendations must be taken in the context of the initial discussions.

Purposes

1 The first task in the setting up of the new survey is to agree formally the purpose(s) of the system. There should be a small number of purposes only and these should be targeted on key areas. The main purposes for a national monitoring survey in England are likely to be:

- monitoring changes to absolute standards over time;
- investigating areas of strength and weakness across the curriculum.

It is not recommended that the tests aim to measure standards in different local authorities, schools or classes, due to the size of the sample that would be required.

2 If it is decided that the purpose of the test is to measure the performance of sub-groups of pupils beyond gender, then the sample must be designed appropriately.

The Sample

3 The size of the sample can only be agreed once decisions about the purposes, any sub-analyses and curriculum coverage are made. It is recommended that research be carried out into the sample size needed once more information is available about the nature of the tests.

4 It is recommended that, for ease of administration and because of cost implications, the basic sample structure of one class per school be considered, rather than small numbers of pupils across a large number of schools. This sample will be used for any written tests, and it is likely that a sub-set of pupils will be used for any additional tests. This will mean that each school will provide a relatively greater proportion of the final data, so the design of the sample will be crucial. Schools should not be identified in any of the results.

The Tests

5 It is recommended that the stakes of the tests be kept low as far as possible, but allowance made for this in the interpretation of the results. Measurement of motivation and attitude to learning and testing should be built into any pilot, and possibly into the final survey design.

6 It is recommended that, although low stakes for the schools, teachers and pupils, participation in the tests should be compulsory.

7 It is recommended that the stratification of the sample be based on school size, school GCSE results, and location of school – rural, urban etc.

8 It is suggested that the tests assess those subjects currently covered by KS3 testing: English (reading and writing), mathematics (including mental maths) and science. The inclusion of ICT in the survey should also be considered.

9 It is recommended that a matrix sampling system is used to assess across the curriculum in depth without overburdening any individual pupils. In this context it is suggested that the assessment of speaking and listening and science practical work be also considered, although budget and manageability constraints may mean that these are not ultimately included. Testing time for pupils should be limited.

10 If the key purpose of the tests is to measure standards over time or performance in different areas of the curriculum and by different sub-groups, then any changes are likely to be small year on year. It is recommended that an initial survey is carried out to set the baseline in all subjects, but that subsequently subjects are assessed on a rolling programme.

11 It is recommended that the use of technology be considered for both the administration of the tests and the marking. It is also recommended that the background data on the pupils and attitudes to learning be collected in the form of an online survey.

Analysis

12 As with many of the decisions to be made on the nature of the tests, decisions about the best means of analysing the data will be dependent on the outcomes of earlier discussions. However, it is recommended that a pragmatic approach be taken to the analysis stage, with an understanding that no one way is the only way to do this, or can provide 'the right answer'. It is recommended that IRT be used alongside professional judgement and classical test theory, to develop the instruments and to draw conclusions about performance in the surveys.

Reporting

13 In terms of areas of the curriculum for reporting purposes, it is recommended that as small a number as possible is chosen to enable the sample size to be kept at a reasonable level. This general reporting can be supplemented by reporting on individual items (for a sample of items, not all of them), and items over time to give more detail. It is recommended that these areas include both content and skill areas.

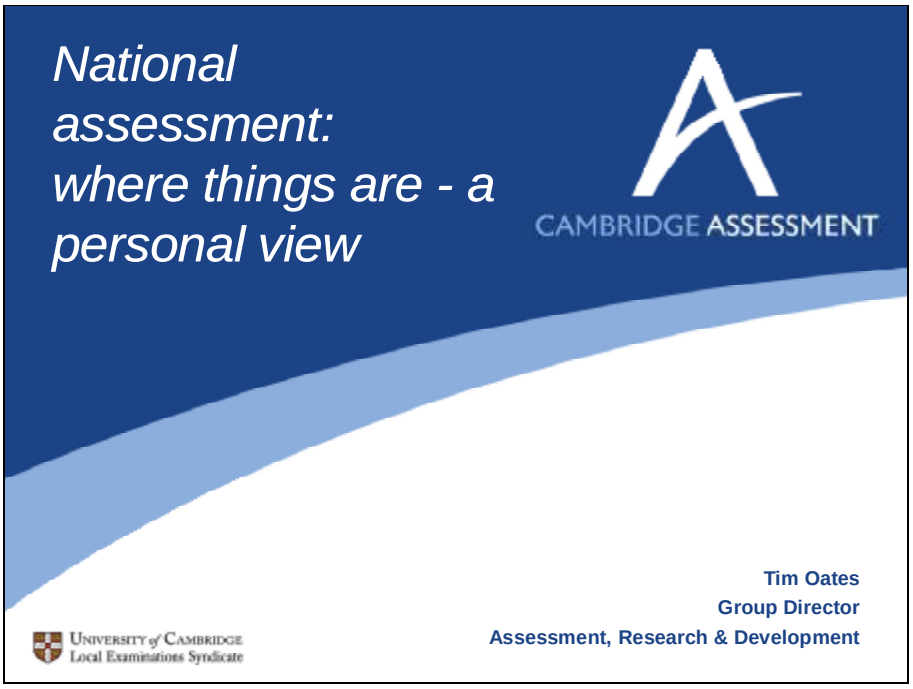
14 Similarly, the number of sub-groups of the population to be reported against should be kept as low as possible to allow the overall sample size to be kept at a reasonable level.

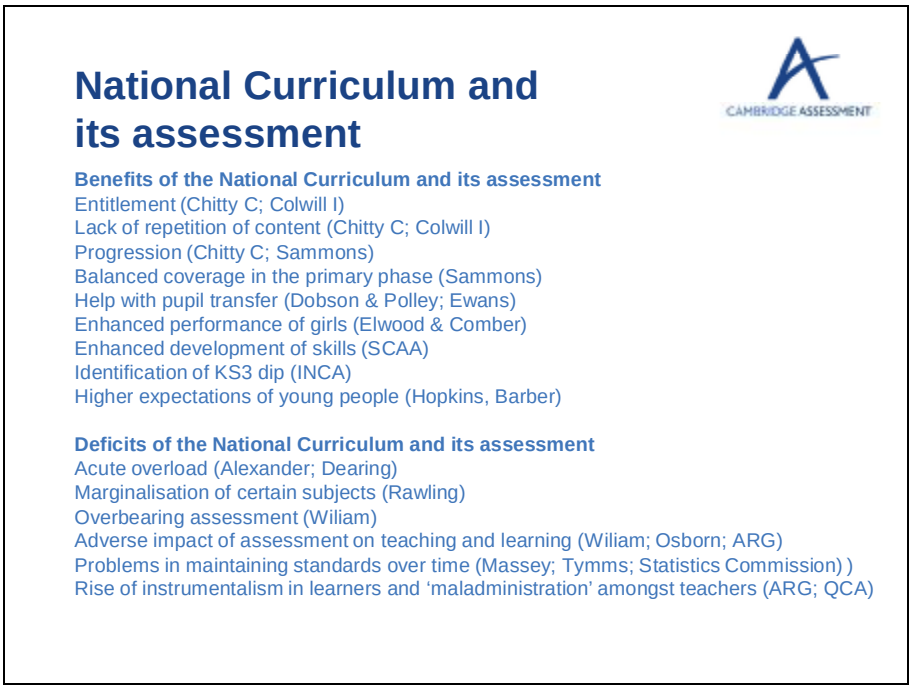
15 The pilot of the new survey should involve the piloting of a 'Nation's Report Card' perhaps with different levels of performance and proportions of pupils at each. These levels should be current National Curriculum levels.

16 It is recommended that the survey design includes ways of linking the results to results from the TIMSS survey, to allow international comparisons to be made.

17 It is recommended that additional in-depth research studies be planned to assess the findings from the main survey in more detail, or to research particular areas of interest at a particular time, rather than trying to cover all the possible needs from the survey in each administration.

Appendix 4: Presentation by Tim Oates





National Curriculum may still need attention

- issues of curriculum coherence (Schmidt and Prawat)
- size and focus (TIMMS, PISA)
- pedagogic deficits (Stigler & Stevenson; Boaler)
- drivers deriving from assessment model (Wiliam; Mansell; Oates, Green, Bell and Bramley)

National Assessment The importance of function

TGAT

1. formative (diagnostic for pupils; diagnostic for teachers)
2. summative (feedback for pupils and parents)
3. evaluative
4. informative

…and in practice: highly intertwined functions

- school accountability
- departmental accountability
- apportionment of funds
- inspection patterns and actions
- upwards pressure on standards/target setting
- structuring of educational markets and school choice
- emphasis of specific curriculum elements and approaches
- detailed tracking of individual attainment, strengths and weaknesses
- quantification of progress

Cambridge Assessment: a revised view of key purposes

National assessment arrangements should provide:

1. formative assessment for teaching and learning
2. information for school accountability
3. information on national standards

Cambridge Assessment and IPPR: alternative models for meeting key purposes

Model 1

Monitoring plus accountability to school level

Model 2

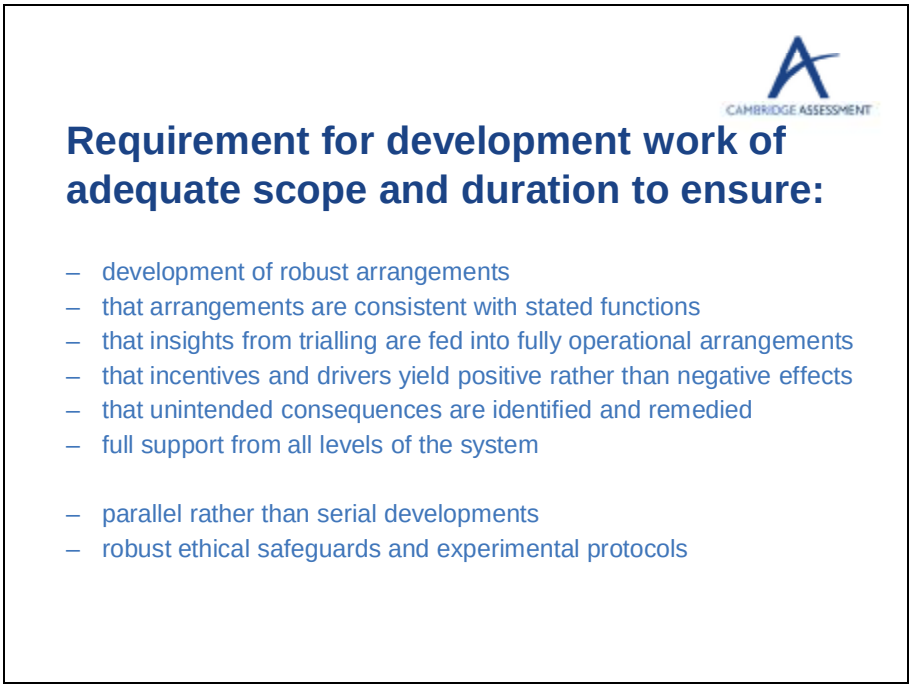
Monitoring plus a switch to 'school improvement inspection'

Model 3

Adaptive, on-demand testing using IT-based tests

The current candidates for filling the void:

- optional use of previous compulsory tests
- the latest incarnation of Single Level Tests
- Assessing Pupil Progress
- commercially-available tests
- commercially-available formative tools



Appendix 5: Presentation by Gordon Stobart

Nuffield Foundation Seminar
National Assessment Arrangements for KS3

Teacher assessment and assessment for learning

Gordon Stobart

Teachers' summative assessment

The process by which teachers gather evidence in a planned and systematic way in order to draw in order to draw out inferences about their students' learning, based on their professional judgment, and to report at a particular time on their student's achievements (ARG, 2006)

It's about purposes

1. Report E/M/Sc the same way as other subjects by teacher assessment – no moderation
2. Move to model with tests/item banks to validate teacher judgments (KS1, Scotland, NZ?)
3. Moderate/ quality assure teacher assessment for local accountability/ transfer. (APP, Wales)
4. Use 3 as basis of KS3 certificate (transfer onto multiple pathways)

Assessment for learning

- It's a distraction? Focus on valid summative assessment of a 'big ideas' curriculum
- Sugaring the pill –making the summative palatable (APP)
- Competing understandings – learning intentions (standards), the inclusion of data (tracking); levelling (MGP) and reliability (APP)
- Day-to-day, Periodic; Transitional assessments