Submission to

Expert Group:
Issues to Consider When Developing
a National Monitoring System

from

National Foundation for Educational Research

Version 2, December 2009

Contents

1 Executive Summary

In October 2008 Ed Balls MP, Secretary of State for Education, announced the end of testing at Key Stage 3 and stated that in its place a system of 'national-level sampling' would be introduced. This paper by the NFER highlights the key issues that should be addressed in the development phase of such a system. The paper provides brief case studies of:

- the Assessment of Performance Unit (APU), the national monitoring system used in England up until 1989,
- the National Assessment of Educational Achievement (NAEP), currently used in the USA,
- the Scottish Survey of Achievement (SSA), the national monitoring system currently in use in Scotland,
- the Trends in International Maths and Science Study (TIMSS), and
- the Programme for International Student Assessment (PISA).

The case studies provide a brief overview of the purpose(s), sample design, the tests, the analysis and the reporting of the systems and a number of issues arising relevant to this development. The main purpose of the case studies is to inform and provide evidence to support the discussion section later in this paper. The main issues are given below.

1.1 Case Studies

1.1.1 APU Key Lessons

1. There is a need for clarity of purpose and definition from the start of the development.
2. Achieving the desired sample proved problematic, especially as the number of subjects and the size of the required sample increased.
3. The administration of the tests with only seven pupils in each school made them logistically difficult to manage.
4. A decision is needed from the start about the reporting that will be required, and therefore the background data that will need to be collected for each pupil.
5. In the APU the monitoring teams were expressly instructed to present only facts in the reports with no interpretation of the findings. This limited the usefulness of the information collected.
6. The method used for analysing the results of the survey was controversial and led to difficulties with reporting certain aspects that were originally required. For a new survey this would need to be addressed early in the process.

1.1.2 NAEP Key Lessons

1. The process of development took much longer than expected and the purpose(s) changed/ evolved. There are regular calls for further expansion of the tests to meet ever more purposes.
2. Measuring change in the system over a long period of time causes challenges, particularly with both keeping the same measure so it can track changes over time, and keeping the measure relevant.
3. There is a disparity of survey results for states with the accountability results of the No Child Left Behind initiative.
4. The system is very complex which leads to misinterpretation of the results in the media and by the public.
5. There are issues with low participation rates among the older students and non-representative samples of students in some sub-groups.
6. The low stakes nature has been linked to some issues with lower than desired response rates and concerns that low motivation may affect the reliability of the results.
7. Even with this low stakes national monitoring system there is still the view that it has led to a narrowing of the curriculum.
8. A key aim of the NAEP tests is to report performance of sub-groups of pupils, such as boys and girls, pupils with disabilities, pupils from different ethnic backgrounds, education of parents and so on. There have been some issues with collecting reliable evidence from the different sub-groups.
9. In the USA item response theory (IRT) is used to analyse the data from the tests, making the assumption that a unitary trait of 'proficiency in the subject' and a 'national population' of pupils are being assessed.

1.1.3 SSA Key Lessons

1. The paper and pencil tests used in each survey spanned the whole curriculum for the subject concerned, with the exception of practical skills. Because of the cost and logistics involved these latter were addressed in a very much less formal, smaller-scale way. Field officers, nominated by their local authorities, conducted and rated the practical assessments. Replicating this type of practical assessment in England could lead to a high-cost system.
2. Items and tasks were 'leveled' (A to F) using professional judgement, on the basis of the 5-14 criterion-referenced progression framework, before being put into the national assessment bank, from which they could be drawn at any time for survey use. In the interests of standardization and interpretability the cut score for 'secure' level attainment on the paper and pencil tests was pre-set at 65% for all surveys and stages.
3. Teachers' level judgements were also collected for the pupils tested in the surveys, although these were not intended for use in system monitoring. Disparities were evident

4. Those teachers who were actively involved in the programme, either as field officers or as raters of pupils' writing, appreciated the professional development experience. This can be seen as a useful additional benefit of the survey programme.
5. There have been some minor concerns about the low stakes nature of the assessment and how this might have affected test performance.
6. Confidence intervals have been reported alongside the attainment results, indicating the precision of the measures being made of population attainment.

1.1.4 TIMSS Key Lessons

1. It frequently proves difficult in England to achieve the required sample of schools willing to participate in the tests. Incentives have recently been introduced.
2. The tests are paper and pencil only and therefore assess a limited proportion of the curriculum.
3. Trends over time as measured by the tests have questionable reliability.
4. The assessment framework reflects the needs of all the participating countries, so does not assess the whole of the National Curriculum.

1.1.5 PISA Key Lessons

1. The assessment of application of knowledge and skills rather than curriculum content is an interesting feature of the PISA surveys.
2. As with other studies mentioned it is not always easy to get sufficient schools to participate. England failed to meet its target in 2003.

1.2 Discussion

1.2.1 Purpose

The first critical decision to be made if a national monitoring system is to be introduced at Key Stage 3 regards the purpose of the assessments. The purpose(s) could include: to monitor standards over time, to monitor performance in different parts of the curriculum or by different groups of pupils, or to develop expertise in the teacher workforce. The decision about the purpose(s) will determine all further decisions about the system, affecting the design of the sample, the design of the tests, the reports that can be produced, and so on.

1.2.2 The Sample

The size of the sample must be chosen to balance the need for precision in the findings at the whole cohort level and for any sub-groups, with the requirement not to over-burden schools. It is likely that some kind of stratified sample will be needed to ensure sufficient representation of different sub-groups and different types of schools.

It is important that the manageability of the tests is also considered, whether this will be with sub-groups of pupils in a school, whole classes or whole year groups. This will affect the amount of disruption in the participating schools. A related issue is that of the stakes of the test, will they be seen as high stakes by the pupils and teachers, or low stakes? High stakes will mean that there is impact on the curriculum that is taught, and stress for those involved, whereas research shows that low stakes can affect performance in the tests due to lack of motivation. As with manageability this issue about stakes will affect decisions in schools about whether to participate or not. In order to ensure the representativeness of the sample it may be desirable to make participation in the survey compulsory, if possible.

The final issue in this area is the level of confidentiality required of the tests. This will be closely related to their stakes. This decision will affect the administration of the tests, eg if administrators are needed to take the materials into schools, and the number of items that can be re-used or released after each administration.

1.2.3 The Tests

An initial decision will be needed regarding the content of the tests, this could be the same subjects and the same aspects of the curriculum as assessed in the Key Stage 3 tests, or could adopt a different approach, such as assessment of more process skills than in the current tests. The breadth of curriculum coverage required will also need to be decided. The tests could be designed to cover the whole curriculum, in which case practical assessments, assessments of speaking and listening etc. will be required. If it is only those aspects that can be assessed via paper and pencil then the design will be much more straightforward and the costs lower, but the results will be less valuable.

The existing surveys described in the case studies all use a matrix sampling approach, that is a large number of items are written across the curriculum, but individual pupils are only presented with a small proportion of these items. A decision is needed as to whether this approach will also be adopted here.

A further issue is the frequency of the testing. This will relate to the purpose, for example if the purpose is to monitor changes to standards over time, these are unlikely to change significantly over one year, so less frequent tests would be appropriate. Finally, a key decision about the tests will be the use of technology in their development, administration and/ or processing. Technology is being much more widely used in teaching and learning and in the administration of tests, and could allow different approaches to be taken to these assessments.

1.2.4 Analysis

How the results from the surveys will be analysed is a key question and will need a significant amount of time set aside to resolve. It is important that a consensus is reached about this in the early stage of the development. This should be seen as a key part of the design phase and not an

afterthought as it will affect the size and design of the sample and the tests. It is likely that analysis will be required at both the pre-test and the live test stages.

There are a variety of approaches that could be used for the analysis, but the most common approach used, as exemplified in the case studies, is item response theory (IRT). NFER uses IRT extensively, but always alongside other means of analysis such as professional judgement and classical test analysis. Each of these sources can provide useful additional information that can contribute to a final decision. It is important to remember that there is no 'right' way; that no one way is perfect, however it is likely that different methods will provide similar results.

1.2.5 Reporting

The level and nature of reporting required will determine the number of pupils in the sample, the number of items, the frequency of the testing, and so on. This is closely linked to the purpose(s) of the system and is one of the first decisions that will need to be made. It is likely that an overall measure of achievement in a subject area will be required, and an overall measure of achievement of the whole cohort. However, the analysis of different topics in each subject, and of different sub-groups of pupils, will impact on design of the system, including the background variables that need to be collected.

Both the purpose and the detail of reporting should be kept as simple as possible as this will impact on the demands placed on schools, the complexity of analysis and the cost of the system. It would be valuable to consider a number of methods that could be used to link the findings from a national monitoring system to results from TIMSS, so that national standards could be linked to those in other countries.

1.3 Recommendations

1.3.1 Purposes

1 The first task in the setting up of the new survey is to agree formally the purpose(s) of the system. 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.

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.

1.3.2 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 is considered, rather than small numbers of pupils across a large number of schools. Schools should not be identified in any of the results.

1.3.3 The Tests

5 It is recommended that the stakes of the tests be kept low.

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.

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 design is used to assess across the curriculum. It is suggested that the assessment of speaking and listening and science practical work be also considered. Testing time for pupils should be limited.

10 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 is considered for the administration of the tests, the marking and the collection of background data.

1.3.4 Analysis

12 It is recommended that IRT is used alongside professional judgement and classical test theory, to develop the instruments and to draw conclusions about performance in the surveys.

1.3.5 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.

14 The number of sub-groups of the population to be reported against should be kept as low as possible.

15 The pilot of the new survey should involve the piloting of a 'Nation's Report Card'.

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 are planned to assess the findings from the main survey in more detail.

2 Introduction

In October 2008 Ed Balls MP, Secretary of State for Education, announced the end of Key Stage 3 testing, and proposed that in its place schools in England should have 'national-level sampling at Key Stage 3'. NFER (National Foundation for Educational Research) is submitting this paper to the expert group on lessons learnt from its experience of assessment development over a number of years, including playing a key role in developing the APU and administering the international surveys (TIMSS, PIRLS and PISA), to inform the proposals for a national monitoring system. Our recommendations are also informed by national monitoring systems in other countries. Details of NFER's credentials are included as Appendix 1.

This paper by the NFER aims to highlight the key issues that need to be addressed when considering the implementation of a national monitoring system. A number of case studies of existing systems (Scotland and USA), from international monitoring surveys (TIMSS and PISA) and from a previous monitoring system in England (APU) have been included to demonstrate the different approaches taken with regard to the issues. A number of recommendations are included, although at this stage of the discussion they must be viewed as possible solutions, as some major decisions need to be made before these can be finalised. The development process for a national monitoring survey is likely to be one of iteration, where initial decisions impact on future discussions and choices.

A list of the documents the NFER hold in our APU archive is also included in Appendix 2 which may be helpful as a reference tool to those deciding on the way forward for a new national monitoring system. The next section gives a number of brief case studies on which the discussion section is based.

2.1 Case Study 1: The Assessment of Performance Unit (APU) in England

The APU was the national monitoring system in England prior to the introduction of the National Curriculum tests in the late 1980s. The APU was first announced in 1974 and a unit was set up in the then DES to run the project. Initially, cross-curricular testing of: verbal, mathematical, scientific, ethical, aesthetic and physical knowledge and skills was considered, but very quickly suites of tests in single subjects were introduced and the first mathematics assessment took place in 1978, followed by language in 1979 and science in 1980. Modern foreign languages (French, German and Spanish) were introduced in 1983 with design and technology following in 1988. An initial purpose of the tests was to investigate underachievement in particular sub-groups of the population but soon after the start of the project the focus changed to assessment and monitoring of performance in different areas of the curriculum, and of different sub-sets of pupils, and to a lesser extent monitoring of standards over time.

2.1.1 The Sample

A sample of 10 000 pupils was chosen for the first maths survey (about 1.5% of the population), initially with the idea of using this large sample as a benchmark and then using smaller samples in subsequent administrations. In reality, once this number was chosen it was carried through to all subsequent surveys. Even with this relatively large number, it was felt that there were difficulties with measuring performance at the extremes of the ability range, one reason why the focus on underachievement was quickly dropped. This was also dropped because there was no agreed definition of underachievement to work to. Only a small number of pupils (usually seven) from each school was included, which made the logistical demands on schools very great. The studies were designed specifically not to be able to identify any individual entities, either pupils, teachers or schools (Foxman, Hutchison and Bloomfield, 1991, p63).

The initial plan was not to administer tests in any school two years in a row, but once the sample size, the number of pupils per school and the full range of subjects were agreed this proved to be impossible to achieve in secondary schools.

Pupils aged 11, 13 and 15 were tested in different subjects and at different times. Different age groups were tested in different subjects.

2.1.2 The Tests

The tests aimed to cover the whole curriculum within each subject area, and therefore included a combination of paper and pencil and practical tasks. For example, in maths there was a written test and a practical test. The tests were designed in such a way that different areas of the subjects could be reported against. In maths the tests included items assessing: geometry, measurement, number, algebra and probability, and statistics. These topic areas were again broken down so that there were 12 or 13 'item clusters' in total that could be reported against. A number of different skills in each subject area were also targeted.

The tests used a matrix sampling approach, that is, a large number of items were developed and these were split over a number of test booklets. The test booklets were divided so that no pupils took all the questions. The large number of questions ensured that the complete curriculum could be assessed. Each test booklet contained two groups of questions and there were overlapping groups of questions between each booklet.

Sub-sets of 2 000 to 3 000 pupils took attitude questionnaires or practical tests in addition to the written tests.

2.1.3 Analysis

The main form of analysis in the language and maths tests was an early form of a one parameter item response theory (IRT) model, known at the time as Rasch analysis. NFER tested out the feasibility of using the Rasch model in a national monitoring program in the TAMS (Tests of Attainment in Mathematics in Schools) project (Sumner, 1975), and then applied it to the APU mathematics survey. This was used as a means of linking items from different test versions within one session, and different tests over time. The APU (English) Language survey did not fit the Rasch model exactly, since item scores were frequently multi-mark, rather than one mark, and the NFER developed extensions to the Rasch model to cope with this. The use of the Rasch model proved controversial, and after some criticism, it was supplemented by other approaches and its use was de-emphasised. The language and mathematics APU studies were discontinued after two rounds. In many ways this analysis methodology was a ground-breaking design but this seems not to have got the degree of credit that it might otherwise have done. As is often the case with leading edge innovation, the difficulties were emphasised and not the positives.

The science tests were developed by a different team, based at Chelsea College/King's College and Leeds University, and analysed using generalizability theory. This was a very different approach to the IRT method. A large 'domain' of items, representing a desired implementation of the curriculum, was created. Samples of items were then selected to form the material for the tests, and these were allocated into booklets according to matrix sampling principles as in other surveys. Overall scores were calculated using a weighted averaging procedure. The APU science survey then used the generalizability theory approach to assess the amount of variability in the overall results due to different aspects of the design. In particular, and differently from other approaches, this allowed an assessment of the degree of variability due to the limited number of items sampled. Unfortunately, producing item pools of sufficient size was a very labour intensive and time consuming process, and the team had only produced a pool of sufficient size when the whole project was discontinued.

There was much debate about the best means of analysis and a key point to note is that in the early 1980s there was no common curriculum in schools in England, and as such there was no guarantee that the pupils had all followed the same curriculum. This caused particular problems for the Rasch model which no longer exist in England's schools today (assuming the target audience is maintained schools). A related issue was the complexity of the data collected and the relative simplicity of the model. Since that time, IRT models that can take into account the effects of a number of different parameters, have been developed and are widely used in national and international surveys around the world.

2.1.4 Reporting

In the APU, reporting was at the regional and the national level. As mentioned above there was no reporting of results for individual schools, teachers or pupils.

Reporting was of different sets of skills or areas of the curriculum, rather than a single overall score (this was in part due to arguments about the unidimensionality[1] of the subjects being measured). Patterns of errors in different clusters of items were reported. The emphasis was on comparing the performance of sub-groups of pupils at particular points in time, so boys and girls, region, school type and so on, or on strengths or weaknesses in different parts of the curriculum, rather than monitoring absolute subject standards over time.

2.1.5 Issues

There were a number of issues with the APU tests which may be of interest to those considering adopting a national monitoring survey.

1. There was a change in purpose as the system developed, affecting the nature of the tests and the reporting that was required. This caused problems for the design of the system. This also led to a growth in the sample size required for each survey, and therefore an increase in the burden on schools.
2. There were difficulties with getting schools to participate which brings into question the generalisability of the results. At the time the difficulties were carefully monitored to ensure they did not impact on the findings. The difficulties experienced at that time were less severe than in the current climate (see the section on TIMSS below).
3. The administration of the tests with only seven pupils in each school made them logistically difficult to manage. This is likely to be linked to issue 2 above.
4. There were issues with collecting background data on the pupils as the consultative committee, lobbied by the teacher unions, continually voted against this. In the end only limited data was available which made it impossible to make as much use of the performance data collected as should have been possible.
5. The monitoring team were expressly instructed to present only facts in the reports with no interpretation of the findings. Again this limited the usefulness of the information collected.
6. A key issue with the success of the APU was the methodology used for its analysis. There was considerable debate about the appropriateness of the Rasch model and

[1] The Rasch model assumes that all the items on a test can be put on to a single scale, that is, they represent a single trait.

whether the underlying assumptions could be met. A discussion of this issue and the different available options is given in the discussion section below.

2.2 Case Study 2: The National Assessment of Educational Progress (NAEP) in the USA

Perhaps the best known of the monitoring systems in other countries is the NAEP in the USA. The NAEP tests were established in 1969 with the aim of authenticating education reforms and to further educational research. Initially one set of tests was set up in reading and mathematics that could be used to monitor attainment over time (the Long-term Trend, or LTT, NAEP). However, information from the tests was required for an increasing number of purposes, and after a successful pilot starting in 1990 a separate suite of tests was introduced that would enable state-by-state comparisons of attainment. This second set of tests became known as the main NAEP tests.

Today the tests serve two main purposes:

- tracking changes in national standards over time;
- comparing achievement across states.

Audiences for the results include: educators, parents, policy makers and the media. No information is provided on individual school or pupil performance. The NAEP runs alongside accountability tests at the state level, used as part of the No Child Left Behind (NCLB) initiative.

2.2.1 The Sample

The main tests are administered in grades 4, 8 and 12 (the equivalent of years 5, 9 and 13), with reading and maths administered every two years and the other subjects less frequently. The LTT tests are used with learners at ages 9, 13 and 17, and are administered every four years.

The main NAEP tests in reading and maths are sat by over 650 000 pupils, a small fraction of the whole cohort of three million students in each grade (in public and private schools). 75 000 students sat for the Long-term Trend NAEP tests in 2004. These latter tests are only used to provide national data.

For the main NAEP approximately 2 500 students from 100 schools are sampled from each state.

2.2.2 The Tests

The tests are only available as paper and pencil tests i.e. no practical assessments are included. Both sets of tests use a combination of multiple choice, short and long answer questions. The tests also both collect a significant amount of background information on the pupils, teachers and schools.

Reading and maths are a compulsory component of both sets of tests. A number of other subjects are available for the main NAEP tests including science, writing, US history, world history, geography, economics, civics, foreign language and the arts. There has been a recent debate about widening the tests to include subjects such as critical thinking, physical and emotional health, work ethic and appreciation of arts and literature, so that purely academic achievement is not the focus. There is a discussion about how far this development of the whole child is the remit of the schools, and a separate movement that believes that education progress should be portrayed by a wide range of factors, of which performance on the NAEP tests is only one factor.

Educational Testing Service (ETS) is the primary NAEP contractor with Pearson and some other organisations playing a smaller role.

Like the APU the NAEP tests again use a matrix sampling approach. To ensure appropriate curriculum coverage hundreds of items are developed for each test. Students are tested in only one subject area and only take a small proportion of the test questions. There are overlapping questions in the different test papers to allow the data to be added together after the testing.

2.2.3 Analysis

All items are pre-tested prior to live usage.

The 1985-86 NAEP was quite similar in general design terms, though not in detail, to the APU tests. Matrix sampling was used, together with a 3-parameter IRT model for analysis. Professional judgements are also used as part of the process, and as with the APU there has been some controversy about the appropriateness of these methods. The scales for each subject are developed independently and therefore cannot be compared e.g. a 250 in reading is not the same as a 250 in maths.

2.2.4 Reporting

Results are given in the form of 'the Nation's Report Card'. Achievement levels in different subjects are reported as: basic, proficient and advanced, so statements such as 50% of grade 8 achieved basic or above, can be made. The particular skills associated with each achievement level are also reported. Success on individual items and what they aimed to assess are also reported. There is no reporting of any form of value added by the schools.

Links have been made from performance in NAEP to performance in international surveys such as TIMSS, although the two are viewed as entirely independent systems.

2.2.4 Issues

1. The process of development took much longer than expected and the purpose changed/ evolved. There are regular calls for further expansion of the tests to meet ever more purposes.

2. Measuring change in the system over a long period of time causes challenges, particularly with both keeping the same measure so it can track changes over time, and keeping the measure relevant.
3. There is a disparity of survey results for states with the accountability results of the NCLB.
4. The system is very complex which leads to misinterpretation of the results in the media and by the public.
5. There are issues with low participation rates among the older students and non-representative samples of students in some sub-groups.
6. The results from the tests are not reported at pupil or school level and there is no reward or penalty for participants, such as pupils, parents, teachers or administrators. This low stakes nature has been linked to some issues with lower than desired response rates and concerns that low motivation may affect the reliability of the results. However, research (Kiplinger and Linn, 1993) has been carried out on the effects of low motivation in the tests compared to higher stakes tests and only small differences have been found, particularly on easy items.
7. A key aim of the NAEP tests is to report performance of sub-groups of pupils, such as boys and girls, pupils with disabilities, pupils from different ethnic backgrounds, education of parents and so on. There have been some issues with collecting reliable evidence from the different sub-groups.
8. In the US IRT is used to analyse the data from the tests, making the assumption that the tests assess a unitary trait of 'proficiency in the subject' and a 'national population' of pupils. There is an on-going debate about the methods used to set the standards and to equate results over time.
9. Even with this low stakes national monitoring system there is still the view that it has led to a narrowing of the curriculum.

2.3 Case Study 3: Scottish Survey of Achievement (SSA)

The Scottish Survey of Achievement (SSA) was introduced in 2005 to replace both the sample-based Assessment of Achievement Programme (AAP), which had been running since 1983, and the separate annual attainment census that was based on submitted teacher judgements for all pupils in all primary stages and in the lower stages of secondary schooling. The SSA aims 'to find out how well pupils are learning in primary schools and the first two years of secondary schooling in Scotland. It is the approach used by the Scottish Government to monitor performance nationally at these stages of pupils' education' (Learning and Teaching Scotland, www.ltscotland.org.uk/assess/of/ssa/introduction).

2.3.1 The Sample

Recent surveys have focused on pupils in primary 3, 5, 7 and secondary 2 (age 8, 10, 12 and 14). At each stage pupil samples have typically been very large, since one of the new purposes of the SSA, over the AAP, was to provide attainment estimates at the level of local authorities as well

as nationally. So that while a typical national sample might be around 4000 pupils at a stage (6-7% of the population), as in 2009, which reported only at national level, the numbers of pupils tested in the surveys of 2005 to 2008 varied between 7,000 and 9,000 per stage. It has not been compulsory for individual schools to take part. Those that do so, and the pupils tested within them, remain anonymous.

2.3.2 The Tests

Subjects assessed in the AAP were English, mathematics and science with social science added in the later stages. These subjects were carried forward to the SSA. When the SSA was first launched the intention was to have annual surveys, each focusing on one or other subject, so that each of the four subjects would be assessed every four years. The choice of pupil stages was such that the same cohort would then be assessed for the same subject at two different points in time (P3 pupils at P7, and P5 pupils at S2). Between 2005 and 2009 English language was assessed twice (2005, 2009), and social subjects, science and mathematics once (in 2006, 2007 and 2008, respectively). Future surveys are to alternate between literacy and numeracy, to relate to Scotland's new Curriculum for Excellence.

Surveys have assessed attainment with reference to the 5–14 progressive level framework (i.e. the whole curriculum). The principal attainment monitoring tool has been paper and pencil testing, with numerous randomly parallel tests being administered to pupils using matrix sampling. Smaller scale exercises have also featured, including in-school practical investigations of various types (administered by a team of field officers) and class-based writing (externally rated). Since the introduction of the SSA, teachers' level judgements have been collected for the same sample of pupils as sit the tests, for research purposes. Background information has also been routinely gathered using pupil and teacher questionnaires.

In mathematics and science the test booklets in 2007 and 2008 contained items at three consecutive levels, with each pupil taking two booklets. Items at the same levels across a pupil's two booklets comprised that pupil's single-level test. In reading and social subjects enquiry skills larger single-level tasks were used.

Those teachers who were involved in the programme, either as field officers or as raters of pupils' writing, felt they had benefitted professionally as a result, and this can be seen as an additional benefit of the survey. The field officers and writing raters were nominated by their local authorities for survey participation.

2.3.3 Analysis

The measurement methodology used to provide the national and local authority attainment estimates on the basis of the pencil and paper testing varied by subject. In reading, where the assessment tasks were relatively long and time-consuming (source text plus sections of related test questions), re-use of a set of the same tasks from one survey to another formed the basis of over-time attainment comparisons. In mathematics, where the pencil and paper tests comprised

relatively atomistic items with no common source materials, stratified domain sampling was employed to select the items for use in a survey. These items were then distributed among a series of randomly parallel single-level tests. The tests were randomly merged to create mixed-level test booklets, and test booklets were assigned at random to individual pupils. The same cut score of 65% correct was used to indicate 'secure' level attainment for all subjects in all surveys, this criterion having been agreed as appropriate by subject specialists in 2001. Attainment was reported as the percentage of pupils achieving the cut-score criterion on the single-level tests. Jackknifing was employed to estimate the standard errors associated with the estimated attainment proportions. Generalizability theory was used in secondary analyses to explore assessment reliability more fully.

2.3.4 Reporting

National reports are available online to all schools each year. As noted above, pupil attainment was reported in terms of the proportions of pupils attaining particular levels in the 5-14 progressive level framework. Confidence intervals are reported alongside the results for the formal national attainment estimates.

The surveys were low stakes with pupils and schools remaining anonymous. The proportions of pupils attaining the levels 'expected' for their stage (the expected levels in the national curriculum guidelines) have been frequently lower than anticipated. This could have been due to the low stakes nature of the testing and a resulting lack of motivation, or to a lack of realism in the 'expected' levels originally set by policy makers.

2.3.5 Issues

1. The paper and pencil tests used in each survey spanned the whole curriculum for the subject concerned, with the exception of practical skills. Because of the cost and logistics involved these latter were addressed in a very much less formal, smaller-scale way. Field officers, nominated by their local authorities, conducted and rated the practical assessments. Replicating this type of practical assessment in England could lead to a high-cost system.
2. Items and tasks were 'leveled' (A to F) using professional judgement, on the basis of the 5-14 criterion-referenced progression framework, before being put into the national assessment bank, from which they could be drawn at any time for survey use. In the interests of standardization and interpretability the cut score for 'secure' level attainment on the paper and pencil tests was pre-set at 65% for all surveys and stages.
3. Teachers' level judgements were also collected for the pupils tested in the surveys, although these were not intended for use in system monitoring. Disparities were evident between the test results and the teacher judgements, especially in science (see Johnson and Munro, 2008).

4. Those teachers who were actively involved in the programme, either as field officers or as raters of pupils' writing, appreciated the professional development experience. This can be seen as a useful additional benefit of the survey programme.
5. There have been some minor concerns about the low stakes nature of the assessment and how this might have affected test performance.
6. Confidence intervals have been reported alongside the attainment results, indicating the precision of the measures being made of population attainment.

2.4 Case Study 4: Trends in International Maths and Science Study (TIMSS)

The TIMSS study is an international survey comparing performance in maths and science across a number of different countries over time. It is conducted by the International Association for the Evaluation of Educational Achievement (IEA) and sits alongside the PIRLS tests which are reading assessments. The survey has been running since 1995 in its current form and England has participated from the start (see Whetton et al 2007 for a review of the studies and performance of primary pupil achievement in them).

2.4.1 The Sample

The sample aims primarily to target students, but also schools and classes. Whole classes are surveyed to make it more manageable. A sample of 4000 students for each subject is used. The sample of schools is randomly selected from all the possible schools in a country, and at the same time two sets of alternative schools, matched school by school, are also selected. If a school declines to take part then the equivalent school from a second matched list is approached, if this school also declines then the school from a third list is approached.

The surveys are kept as brief as possible to minimise the assessment burden, and unreleased questions are used in later surveys to allow linking and scaling, and long term tracking of responses.

TIMSS assesses pupils in the equivalent years to US grade 4 and grade 8, so in England this is year 5 and year 9. After difficulties achieving the sample in 2003, in 2007 the stringent sampling criteria were comfortably met. At grade 4, aiming for a sample of 160 schools in England, 131 from the first list participated for the year 5 tests, with 12 from the replacement lists, giving a total of 143 schools in the sample. For grade 8 (year 9), again a target of 160 schools was required and 126 from the initial list participated, with 11 replacement schools, giving a total of 137 participating schools.

A number of background variables are collected for the students to inform the analysis:

1. Curriculum questionnaires address issues of system-wide curriculum design and support, and curricular emphasis on maths and science;

2. A school questionnaire asks school leaders to provide information about the major factors affecting student success in maths and science;
3. Teacher questionnaires asked maths and science teachers about their preparation to teach, their teaching activities and approaches, their attitudes towards teaching the subject matter, and the curriculum that is implemented in the classroom;
4. A questionnaire for students seeks information about their home backgrounds and resources for learning, their attitudes towards maths and science, and their experiences in learning these subjects.

2.4.2 The Tests

The test development cycle runs for about two and a half years, involving subject specialists and test development experts from around the world. Initially the assessment frameworks are updated to reflect any changes to the curriculum followed in the different countries. A large number of items are then written for expert review and pre-testing.

The framework has two dimensions: content and cognitive (e.g. in maths in 2007 for grade 4 the content dimension was number, geometric shapes and measures, and for cognitive domains this was data display, and knowing, applying and reasoning). Calculators are optional in grade 8 and countries can choose to use them if this best reflects how the children are taught. Calculators are not permitted in grade 4.

Again using 2007 as an example, 353 items were used in the 4^{th} grade maths tests and 429 items for 8^{th} grade. The items were separated into 14 blocks at each grade – and combined with 14 for science. The blocks each take 18 minutes to complete for 4^{th} grade and 22.5 minutes for 8^{th} grade. Some blocks contained secure items from earlier tests. Each final booklet contained four blocks, two science and two maths blocks. The blocks were distributed across 14 student booklets. To enable linking each block appeared in two booklets. In total the time allocated was 72 minutes for 4^{th} grade and 90 minutes for 8^{th}. A balance across blocks and booklets for content and cognitive areas was aimed for where possible. Unreleased items will be used in future tests to link standards over time.

2.4.3 Analysis

The Third International Mathematics and Science Survey (TIMSS) took over some of the personnel, and consequently some of the methods, used in NAEP. The first round of TIMSS was analysed using the one parameter IRT method, but later rounds of TIMSS used the three parameter IRT method. The 1995 study, originally analysed using the one parameter method, was later re-analysed using the three parameter method.

2.4.4 Issues

1. It frequently proves difficult in England to achieve the desired sample of schools willing to participate in the tests. For 2007 an incentive was introduced to encourage schools to participate which has had the desired outcome.
2. The tests are paper and pencil only and therefore assess a limited proportion of the curriculum.
3. Trends over time as measured by the tests have questionable reliability.
4. The assessment framework reflects the needs of all the participating countries, so does not assess the whole of the National Curriculum.

2.5 Case Study 5: PISA

Much of the early development work in connection with PISA actually took place within the IEA setup, that is within the organisation which administers TIMSS. For this reason, while the actual aim and philosophy of testing differ between PISA and TIMSS, it is not surprising to find that survey design and data analysis methods do not differ largely. Both studies used a matrix sampling design and IRT methods.

However there are a number of ways in which PISA differs from other surveys which may be worth highlighting here.

Firstly, PISA does not assess traditional educational curriculum areas. PISA is conducted by the OECD (the Organisation for Economic Cooperation and Development) and is reactive to the desires of national policy makers, as such they are interested in the skills required to support a successful economy. The study targets pupils at the school leaving age, and skills such as mathematical literacy, scientific literacy and reading literacy, rather than curriculum-based content.

Secondly, the programme has surveys every three years, with each subject included each time but only one as the 'main focus', which means that each subject is only studied in depth once every nine years. In 2000 the main focus was reading literacy, in 2003 it was numeracy and in 2006 it was scientific literacy. Finally, PISA uses a one parameter model, described as a Rasch model, in analysing the results.

2.5.1 Issues

1. The assessment of application of knowledge and skills rather than curriculum content is an interesting feature of the PISA studies.
2. As with other studies mentioned it is not always easy to get sufficient schools to participate. England failed to meet its target in 2003.

3 Discussion

The previous national monitoring system in England (APU) began in 1974 and came to an end with the introduction of the National Curriculum tests in 1989. There were a number of issues with the APU related to the methodologies used for analysing the results and the issues associated with monitoring standards over time, as well as the collection of background variables and the usefulness of the results. Over the last few years a number of changes have occurred that make some of the difficulties less severe now and the change in educational culture may mean that a national monitoring system is something that is now more positively viewed. The most significant cultural change has been the introduction of the National Curriculum and its associated testing. At the time of the APU there was no agreed curriculum in England and local authorities and schools had greater freedom about what was taught and when. Criticisms levelled at the APU in terms of it exerting too great an impact on the curriculum and causing a narrowing of teaching, have been superseded by an even greater impact on the curriculum as a result of the National Curriculum tests, and the accountability associated with the results. It is likely that the proposed monitoring tests will now feel 'light touch' and far less intrusive than they appeared in 1974 when they were originally announced.

Over the same period of time the NAEP tests in the USA and the national monitoring tests in Scotland have continued and evolved, and although not without opponents, have demonstrated that a national monitoring system can provide useful information about pupil performance over a period of time. There are a number of useful lessons that can be learnt from a study of NAEP and the Scottish tests, the information they provide, and the methodologies used, set out in the discussion below.

Similarly, the continued use of the TIMSS studies and the introduction of PISA mean that there are a number of established surveys in England through which expertise in sampling, data collection and analysis has been developed. The results from these studies are also likely to contribute useful information when considering a methodology for the new system, and also provide an international context against which results from a new monitoring system can be set.

There are many issues which will need to be thoroughly discussed prior to the introduction of the new system, and it is essential that sufficient time is given over to the development and piloting of the new assessments. The following discussion gives an overview, although not an exhaustive list, of some of the key issues.

3.1 Purpose(s)

An initial key question in the introduction of a national monitoring system is the purpose of the survey itself and the uses that will be made of the results. In the APU the monitoring of standards over time was seen as secondary to the measurement of strengths and weaknesses

within subjects and in different sub-sets of the population. The PISA tests have been introduced to assess the readiness of the respective school populations for employment, rather than focusing on achievement in the educational curriculum and therefore assess literacies, rather than curriculum content. In the NAEP tests a wide number of subjects are included and there is discussion about the inclusion of non-academic subjects in future years. Also in NAEP, performance at the item level is reported, as well as proportions of pupils reaching different levels of achievement, demonstrating clearly that the detail and nature of the agreed purposes and therefore the reporting needed will have a significant impact on the design of the tests. Finally, in the SSA one additional benefit is the development of assessment expertise in the teacher workforce, leading to the inclusion of teachers in the administration and marking of the tests.

3.2 The Sample

The size of the sample must be chosen to balance the need for precision in the findings at the whole cohort level and at the level of any sub-tests or sub-groups of pupils, with the requirement not to over-burden schools. This is a relatively complex decision depending on the number of subjects being monitored, the frequency of the surveys and the need for analysis of sub-sets of the data. For example, an issue with the APU was the difficulty of assessing the extent of underachievement because the sample size and structure did not allow for accurate measurement at the extremes of performance.

Rather than randomly selecting schools it is possible that a stratified sample could be chosen to ensure that different sub-groups are sufficiently represented, and to ensure that the sample is representative of the whole population. A number of different factors can be selected for stratifying the sample, including:

- School size;
- School type: urban, suburban, rural;
- Type of local authority (London borough, metropolitan, unitary and county);
- Government office region;
- Overall level of achievement (using appropriate assessment results);
- Level of disadvantage (as indicated by the percentage of pupils known to be entitled to free school meals).

The stratifying factors would be selected based on the requirements of the final reports. In our experience three stratifying variables are usually sufficient in the English context. Again the number of factors will depend to some extent on the aims of the assessment programme, the size of the sample, and more importantly on the analysis that will be required of the results.

An issue relating to the sample size and the manageability of the tests is the number of pupils assessed per school. In the APU only a small number of pupils per school were assessed, whilst

current studies such as TIMSS assess all the pupils in a class, making the administration much more straightforward for the schools involved.

Another important issue to consider is the 'stakes' given to the tests. The tests could be used to provide information about national performance over time, and if required, performance by local authority (although this would need a very large sample). It is probably not possible to use a sampling test to obtain measures of individual teachers or pupils, and as such it will be possible for the tests to be administered in a low stakes environment. There is a lot of evidence that performance on low stakes tests is significantly different to performance on high stakes tests (Weiss and De Mars, 2005), and more importantly it may be that the effect is more complicated than just an underestimation of overall performance (Pyle et al, 2009).

Related to this point is the issue of whether participation in the tests is optional or not. In the APU there were some difficulties in getting sufficient schools to participate to enable the agencies to be confident that the results were really representative of performance of the whole population. As new subjects were added to the survey, non-response rates increased to a maximum of 23% in the science tests in 1980. In NAEP and in the SSA there have been issues with non-participation. Similarly in more recent times in England it has been difficult to achieve the required samples in TIMSS and PISA (so that England did not fully achieve its grade 8 sample in 2003 in TIMSS and similarly in PISA in 2003 England did not achieve its sample). This has been viewed as an important enough issue to introduce an incentive package to encourage schools to participate. Both in the later years of the APU and currently in the SSA there have been discussions about the information that can be returned to participating schools as a means of encouraging schools to take part. If the tests are seen as low stakes and optional, it may be that it is difficult to achieve the desired sample, and it may also be that pupil performance is affected by bias caused by lack of motivation.

A further related issue dependent on the stakes of the tests is the level of security required and the option to re-use items for equating purposes. In most of the case study examples given above a number of items are released each year, whilst others are kept secure and re-used in future years' tests for equating purposes. Once tests become high stakes it becomes more difficult to keep a number of items secure without large teams of administrators taking the tests into schools, and without the costs associated with this approach.

3.3 The Tests

An initial decision will obviously be needed on the subjects to be assessed as part of the monitoring survey. The Key Stage 3 tests currently assess pupils in English, mathematics and science. The APU tests assessed pupils in mathematics, science, language, modern languages and design and technology, and the SSA and NAEP assess different combinations of subjects again. Obviously the choice of the subjects gives out a message as to what is valued within the curriculum, but the subjects to be assessed may be affected by a curriculum backwash as part of the assessment process.

The PISA tests focus on literacy, mathematical literacy and scientific literacy, thereby giving the idea that it is being able to use the subject knowledge and skills that is most important. This may reflect current initiatives in England such as the introduction of the functional skills tests. An original aim for the APU tests was to assess cross-curricular areas, rather than single subjects.

The subjects being selected for assessment will also impact on the different assessment methodologies being used. For example an English assessment aiming to cover the whole curriculum requires the inclusion of assessments in speaking and listening as well as reading and writing. The APU tests were innovative in their use of practical assessments and included assessments such as speaking and listening in the language tests, and practical work in the maths tests. The development of these methodologies then went on to inform the future development of practical and coursework assessments used as part of the GCSEs and the original National Curriculum tests. In the Key Stage 2 and 3 tests mental mathematics is assessed using audio tapes. The SSA aims to assess across the whole curriculum, but NAEP limits itself to paper and pencil tests. A decision will need to be made about how far these alternative methodologies will be used in the new surveys, and again this will be a balance between manageability and cost.

The number of items is another key issue that needs to be considered. The numeracy section of the Scottish Survey of Achievement (Scottish Government, 2006) may be taken as a fairly typical example of a matrix sampling approach. It was agreed that items would be set across the whole curriculum rather than sampling particular areas and then generalising from performance in those to the whole curriculum. Tasks were randomly allocated into 'booklets' to meet a given booklet specification. At each stage 12 different booklets were prepared. The 12 booklets were paired following an incomplete block design. Each pupil took a booklet pair, allocated at random, so that every booklet was eventually attempted by similar numbers of pupils in similarly representative pupil subsamples. In any one school at most two pupils would attempt the same booklet.

TIMMS, NAEP, SSA and the APU tests all use a matrix approach to curriculum coverage. However, the use of the matrix model of assessment does impact on the size of the sample required for the survey, and also impacts on the complexity of the analysis, as it is necessary to combine the results from the different tests and different pupils back into a single measure of achievement against the curriculum.

A related issue is the types of questions being included in the written tests. In the current National Curriculum tests there is a combination of objective and short answer questions in the maths and science tests, both of which tend to be fairly easy to mark reliably (provided there is an appropriate programme of marker training). The English tests call for longer responses eg to assess writing, which take longer to mark and cause more issues in terms of inter-marker reliability. However, in order to have a valid assessment of English it is likely for this to be necessary. The availability of markers and the demands placed on the system by the administration of the tests will need to be considered when the subjects and focus are decided.

A further decision that will need to be made is the frequency of the test administration in each subject. In the SSA there is a four year rolling programme and a similar approach was adopted in the later years of the APU. The TIMSS tests are administered every four years. In PISA all subjects are tested on each administration (every three years) but there is a main subject which is investigated in detail, and subsidiary subjects.

The timing of the tests will also need to be agreed. In TIMSS the tests for year 9 pupils are about 70 minutes long, however, at Key Stage 3 the pupils each took three subjects so, in fact, were subjected to 7 hours and 35 minutes of testing. Again this is a balance between the assessment burden placed on any individual pupils, the number of questions being included in each test booklet, and the size of the required sample.

A final issue to be considered regarding the tests is whether and how technology is to be used in the assessment process. None of the case studies use technology in the administration of the tests or the surveys but this may be a factor more of when they were introduced than what may be possible now. The international surveys are moving slowly to computer administration and there has been some research into the computer delivery of aspects of NAEP. Although the delivery of tests on computer is still not widely used in England, there is widespread use of scanning and online marking at GCSE in particular, and extensive use of online surveys and analysis. The use of such approaches may make it possible to include more complex assessments or larger samples if they could be shown to lead to time savings in the processes.

3.4 Analysis

In much educational research, details of sample design and methods of analysis tend to be something of an afterthought. The basic methods and approaches are considered to be reasonably well known and discussions tend to concentrate on sample size and arrangements required. In studies such as these, methods, starting from APU and NAEP, were innovative, central to the project and strongly influenced the way the sample was designed.

An initial decision that is needed is the stages at which analysis will take place. The existing Key Stage 3 tests are pre-tested prior to their use, as are the TIMSS and NAEP items. It is likely that there will be a need to pre-test the items for the new monitoring survey and a decision will need to be made about the role of the pre-testing (what information is required from it), the timing of the pre-tests – will they be administered one year ahead with the cohort of the same age, as is the current approach in National Curriculum test developments? Will the pre-tests be used to select the best performing items or to contribute to the standards setting process?

The type of larger scale studies described in this paper have different aims from the more widely familiar pupil testing programs such as GCSE or National Curriculum assessment. Consequently they have to be designed to fulfill different criteria. The next topic to consider is how to analyse, summarise and report such results. The NAEP survey was originally aimed to report percent correct at an item level, but it was soon decided that some level of aggregation was need to communicate results effectively.

There are a variety of approaches, of differing degrees of complexity, to how to obtain such results. We now list them briefly, and describe them in more detail below.

1. Separate test programmes for each topic area with no matrix sampling.

The remainder of the approaches described here use some kind of matrix sampling approach.

2. Simple score aggregation
3. Item Response Models (Rasch and IRT)
4. Generalizability Theory.

1. **Separate programs with no matrix sampling.** One could simply produce a separate test for each area of the curriculum considered and administer it to a large randomly-drawn sample of pupils. However, it is generally accepted that a matrix sampling approach will allow for more of the domain to be tested and will be more efficient in a statistical sense because of the degree of overlap and correlation between tests in the design.
2. **Simple score aggregation.** The Scottish Survey of Achievement (Scottish Government, 2006) used a matrix sampling approach for numeracy with each pupil being presented with two booklets. Each booklet pair contained enough items to classify pupils using three different single level tests. The proportions of pupils classified into the three levels were calculated separately for each booklet pair and the resultant proportions were weighted and averaged to produce population estimates.
3. **Item Response Models (Rasch and IRT).** These are essentially factor analysis methods for one mark or multi-mark data which take performance on a series of scored items as indicators of some underlying trait of 'ability' or 'attainment'. The Rasch model and the one parameter logistic model are essentially the same and assume that the behaviour of any item in a scale can be summarised by a single difficulty parameter. IRT programs will readily produce pupil 'ability' estimates on a comparable scale if each pupil takes the same items, or an overlapping sample of items. At the same time they will produce 'difficulty' estimates for the items.
   In the studies described here, items are distributed among booklets, and these are 'linked' by having pairs of booklets taken by the same pupils. IRT programs can easily cope with this type of arrangement, provided there is this 'linking' and the assumptions underlying the model hold[2]. This means that it is possible to assess pupil performance and item parameters on a common scale even though pupils take different tests. Similarly, it will be possible to use a comparable linking structure by repeating booklets from one year to another, allowing comparisons to be made over time. Provided the model continues to hold over the interval, it permits the programme to replace items that have been made publicly available, and continue to make measurements on the same scale.

[2] The main assumptions are the unidimensionality of the test content and of the population being assessed.

4 **Generalizability theory.** Generalizability theory is a development from Classical Test Theory, and breaks down variance in a measurement into components arising from different sources of variation. Thus for example components of variation in a national average test score could be differences between pupils, between schools, between items and otherwise unexplained. As earlier, the score is determined by simple (possibly weighted) aggregation and averaging: the main strength lies in assessing the contribution of different sources of variation. In some ways it is almost more appropriate as a tool for designing a study than for analysing one.

The controversy over the use of IRT methods has died down to a large extent since their use in the APU. Even in the UK, IRT methods are now used quite widely. Nevertheless many of the original objections were theoretically plausible, and it would be important to set aside time for methodological development and validation research in any such programme to agree on the best means of analysing the test data.

At the NFER IRT techniques are always used alongside additional information in the form of classical test analysis and professional judgements. Each source of information can provide useful additional information that can contribute to the final judgements about performance, and it is important to remember that equating of data between tests or over time is not clear cut, and that no one method is likely to be the one way to do the analysis.

3.5 Reporting

The level and nature of reporting required will, in part, determine the number of pupils in the sample, the number of items, the frequency of the tests, the design of the sample, and so on. This is one of the first decisions that needs to be made as it will impact on many other decisions. It is likely that overall performance in the different subjects will be required (although in the APU this was not a key focus). However it may be that this information is not required annually as it is unlikely that small changes will be detectable on an annual basis, so perhaps the subjects could be administered on a rolling programme, every three or four years, as in the SSA, TIMSS or PISA.

Also, what sub-sets of performance will be required? It is likely that information will be required for sub-sets of pupils, such as by gender and by region. In addition, results may be required by local authority, for different social economic status or for different ethnic groups. This decision was very problematic in the APU and extensive discussions took place about which background variables ought to be collected. In the end very few variables were collected limiting the extent of reporting that was possible.

It is likely that sub-groups of items will also require reporting on. This could be at the topic level, such as number, geometric shapes and measures in TIMSS maths, or by skill such as data display, and knowing, applying and reasoning, also from TIMSS. Finally, a decision will need to be made about whether there will be a need to report at the item level as in NAEP. This will

mean that the items will need to be published for the results to be meaningful, so it would not be possible to keep some items secure as in TIMSS (unless only some items are reported on).

In the APU a decision was made to keep the main survey as simple as possible, thereby reducing the size of the required sample and tests, with the idea of running additional in-depth surveys to explore certain areas in more detail. In actuality these 'in-depth surveys' did not take place, although this remains a useful option when considering the reporting that will be required. In the NAEP, two different sets of tests are used to provide the different types of reports that are required.

There is a movement in the UK to report confidence levels alongside tests results, as is the case in the US. Confidence levels are reported alongside the SSA results. It should be decided whether these should be reported with the results of a new national monitoring system.

A further aspect to the possible reporting would be to link findings from a national survey to international standards through incorporating TIMSS items or similar. This would allow the results from England to be directly compared to the countries included in TIMSS. By using a wider set of questions rather than just TIMMS items it will also allow for full coverage of the National Curriculum.

4 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.

4.1 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.

4.2 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.

4.3 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.

4.4 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.

4.5 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.

5. References

Foxman, D., Hutchison, D., & Bloomfield, B. (1991). *The APU experience 1977-90*. London: SEAC.

Johnson, S. and Munro, L. (2008). Teacher judgement and test results: should teachers and tests agree? Paper presented at the annual conference of the Association for Educational Assessment - Europe, Hissar, Bulgaria.

Pyle, K., Jones, E., Williams, C., & Morrison, J. (2009). *Investigation of the factors affecting the pre-test effect in national curriculum science assessment development in England.* Educational Research 51 (2)

Scottish Government (2006). *Scottish Survey of Achievement: practitioner's report.* Edinburgh: Scottish Government.

Sumner, R. (1975). *Tests of attainment of mathematics in schools: monitoring feasibility study*. Slough: NFER.

Kiplinger, V. L. and Linn, R. L, (1993). *Raising the Stakes of Test Administration: The Impact on Student Performance on NAEP*. Los Angeles, C.A.: National Centre for Research on Evaluation, Standards and Student Testing (CRESST).

Whetton, C., Ruddock, G. And Twist, L. (2007). *Standards in English Primary Education: the International Evidence.* Primary Review Survey 4/2, University of Cambridge Faculty of Education.

Wise, S.L., and De Mars, C. E.(2005). *Low examinee effort in low stakes assessment: problems and potential solutions.* Educational Assessment, 10, 1, 1-17.

Appendix 1: NFER Credentials

The National Foundation for Educational Research (NFER) was founded in 1946, and is Britain's leading independent educational research institution. It is a charitable body undertaking research and development projects on issues of current interest in all sectors of education and training. The Foundation's mission is to gather, analyse and disseminate research based information with a view to improving education and training. Its membership includes all the local authorities in England and Wales, the main teachers' associations and a large number of other major organisations with educational interests, including examining bodies. It is overseen by a Board of Trustees.

The NFER's Department for Research in Assessment and Measurement is one of two research departments of the Foundation. It specialises in test development and research into assessment-related questions. The work of the Department involves projects of importance to national educational policy and its implementation through research, the development of assessment instruments and the evaluation of assessment initiatives. It has a consistent track record of developing high quality assessment materials to meet the needs of a variety of sponsors. The Department's experience covers the whole range of tests and other assessments. NFER's work in assessment and surveys stretches back over its entire history, such that the Foundation has a unique experience of test development and the use of tests. In addition to developing assessments, we also carry out major evaluation studies, large scale surveys and international surveys for a number of sponsors including: DCFS, QCA, Scottish Government and DCELLS.

Experience in Assessment

The following list of projects illustrates the variety of experience in assessment matters:

National Assessment by Sampling the Cohort

NFER was responsible for the greater part of the work of the Assessment of Performance Unit (APU) in the UK. National monitoring of performance in mathematics, English and foreign

language, in England, Wales and Northern Ireland, was undertaken by the Foundation from the early 1970s to the late 1980s, when National Curriculum tests replaced a sampling approach.

National Assessment by Testing the Whole Cohort

Since 1989, the Foundation has undertaken much work in producing National Curriculum tests to be used by the whole cohort in England. Such work has encompassed English, mathematics and science for various ages: 7, 11, and 14 and has been undertaken under contract to QCA or its predecessors. Each of these tests is taken by 600,000 students, and the results have high stakes for schools since they are published as part of the accountability of the education system.

UK Assessment in the International Context

The Foundation has had a long involvement with international assessment, and was a founder member of the International Association for the Evaluation of Educational Achievement (IEA), which was set up in the 1960s and organises international comparative studies of educational achievement. NFER has been responsible for managing the testing for all of the IEA surveys in which England has participated, including both TIMSS (Trends in International Mathematics and Science Surveys) and PIRLS (Progress in International Reading Literacy Survey).

In 2005, NFER became responsible for the OECD PISA (Programme For International Student Assessment) surveys in England, Wales and Northern Ireland for 2006, which will report this year and will also be undertaking the 2009 surveys in all four UK countries.

Appendix 2: List of APU Publications Available

Archenhold, F. (Ed) (1988). *Science at Age 15: a Review of APU Survey Findings 1980-84.* London: HMSO.

Archenhold, F., Austin, R., Bell, J., Black, P., Bround, N., Daniels, F., Holding, B., Russell, A. and Strang, J. (1991). *Profiles and Progression in Science Exploration* (Assessment Matters No 5). London: SEAC.

Assessment of Performance Unit (1978). *Language Performance*. London: DES.

Assessment of Performance Unit (1978). *Monitoring Mathematics*. London: DES.

Assessment of Performance Unit (1979). *Science Progress Report 1977-78*. London: DES.

Assessment of Performance Unit (1979). *Science Progress Report 1977-8 Appendix: List of Science Concepts and Knowledge*. London: DES.

Assessment of Performance Unit (1980). *Extracts from the Report of the West Indian Study Group*. London: DES.

Assessment of Performance Unit (1980). *Foreign Language*. London: DES.

Assessment of Performance Unit (1981). *Personal and Social Development*. London: DES.

Assessment of Performance Unit (1981). *Understanding Design and Technology*. London: DES.

Assessment of Performance Unit (1982). *The Language Performance in English of Fifteen-Year-Old Pupils in Wales: Supplement to the Report on the Second APU Secondary Survey*. Cardiff: Welsh Office.

Assessment of Performance Unit (1983). *Aesthetic Development*. London: APU.

Assessment of Performance Unit (1983). *Exposure and Performance: an Investigation of Test Validity within the Context of the APU Monitoring Programme. Summary Report*. London: APU.

Assessment of Performance Unit (1983). *Foreign Language Provision: Survey of Schools* (APU Occasional Paper 2). London: DES.

Assessment of Performance Unit (1983). *How Well Can 15 Year Olds Write?* (APU Survey). London: DES.

Assessment of Performance Unit (1983). *Physical Development*. London: DES.

Assessment of Performance Unit (1983). *Report of Survey of Design and Technological Activities in the School Curriculum*. Nottingham: National Centre for School Technology.

Assessment of Performance Unit (1983). *Report of Survey of Design and Technological Activities in the School Curriculum: Part 2*. Nottingham: National Centre for School Technology.

Assessment of Performance Unit (1985). *Oracy: Practical Assessment Age 11* (Video). London: Philip Harris Biological.

Assessment of Performance Unit (1985). *Sample Questions in Science at Age 11*. London: DES.

Assessment of Performance Unit (1985). *Science at Age 11 and 15: Sample Questions*. London: DES.

Assessment of Performance Unit (1985). *Science in Schools Age 11: Report No 4*. London: DES.

Assessment of Performance Unit (1985). *Science: Practical Assessment Age 11* (Video). London: Philip Harris Biological.

Assessment of Performance Unit (1985). *Science: Practical Assessment Age 15* (Video). London: Philip Harris Biological.

Austin, R., Holding, B., Bell, J. and Daniels, F. (1991). *Patterns and Relationships in Schools Science: a Booklet for Teachers* (Assessment Matters No 7). London: SEAC.

Black, P., Harlen, W. and Orgee, T. (1984). *Standards of Performance- Expectations and Reality: a Study of the Problem of Interpreting the APU Science Surveys*. London: DES.

Boyce, C. and Portal, M. (1987). *Foreign Languages Listening and Reading: Implications of the APU Surveys for Teaching and Learning*. Windsor: NFER-NELSON.

Brooks, G. (1987). *Speaking and Listening: Assessment at Age 15* (APU Survey). Windsor: NFER-NELSON.

Brooks, G., Cato, V., Fernandes, C., Gorman, T., Kispal, A. and Orr, G. (1997). *Reading Standards in Northern Ireland in 1996*. Slough: NFER.

Burstall, C. and Kay, B. (1978). *Assessment: the American Experience*. London: DES.

Cambridge Institute of Education (1985). *New Perspectives on the Mathematics Curriculum: an Independent Appraisal of the Outcomes of APU Mathematics Testing 1978-82*. London: HMSO.

Department for Education and Employment, Quality and Financial Assurance Division and Performance Improvement Dissemination Unit (2000). *Delivery of Key Skills in Modern Apprenticeships* (QPID Study Report No. 89). London: DfEE.

Dickson, P. (1986). *Assessing Foreign Languages: the French, German and Spanish Tests*. Windsor: NFER-NELSON.

Dickson, P. (1987). *Foreign Languages Speaking: Implications of the APU Surveys for Teaching and Learning.* Windsor: NFER-NELSON.

Dickson, P., Boyce, C., Lee, B., Portal, M. and Smith, M. (1985). *Foreign Language Performance in Schools: Report on 1983 Survey of French, German and Spanish*. London: DES.

Dickson, P., Boyce, C., Lee, B., Portal, M. and Smith, M. (1986). *Foreign Language Performance in Schools: Report on 1984 Survey of French*. London: DES.

Dickson, P., Boyce, C., Lee, B., Portal, M. and Smith, M. with Kendall, L. (1987). *Foreign Language Performance in Schools: Report on 1985 Survey of French*. London: HMSO.

Donnelly, J. (1988). *Metals at Age 15: Responses by 15-Year-Olds Pupils to APU Questions Based on the Properties of Metals*. London: DES.

Driver, R., Child, D., Gott, R., Head, J., Johnson, S., Worsley, C. and Wylie, F. (1984). *Science in Schools Age 15: Report No 2*. London: HMSO.

Driver, R., Gott, R., Johnson, S., Worsley, C. and Wylie, F. (1982). *Science in Schools Age 15: Report No 1*. London: HMSO.

Eggleston, S.J. (1983). *Learning Mathematics: How the Work of the Assessment of Performance Unit Can Help Teachers* (APU Occasional Paper No. 1). London: DES.

Foxman, D. (1987). *Assessing Practical Mathematics in Secondary Schools*. Windsor: NFER-NELSON.

Foxman, D. (1992). *Learning Mathematics & Science: the Second International Assessment of Educational Progress in England*. Slough: NFER.

Foxman, D. (1997). *Educational League Tables: for Promotion or Relegation? A Review of the Issues*. London: ATL.

Foxman, D. (1999). *Mathematics Textbooks Across the World: Some Evidence from the Third International Mathematics and Science Study (TIMSS)*. Slough: NFER.

Foxman, D., Hutchison, D. and Bloomfield, B. (1991). *The APU Experience 1977-1990*. London: SEAC.

Foxman, D., Martini, R.M. and Mitchell, P. (1982). *Mathematical Development Secondary Survey Report No 3*. London: HMSO.

Foxman, D., Ruddock, G. and McCallum, I. (1990). *APU Mathematics Monitoring 1984-88 (Phase 2): a Summary of Findings, Conclusions and Implications* (Assessment Matters: No. 3). London: SEAC.

Foxman, D., Ruddock, G. and Thorpe, J. (1989). *Graduated Tests in Mathematics: a Study of Lower Attaining Pupils in Secondary Schools*. Windsor: NFER-NELSON.

Foxman, D., Ruddock, G., Joffe, L., Mason, K., Mitchell, P. and Sexton, B. (1985). *Mathematical Development: a Review of Monitoring in Mathematics 1978 to 1982. Part 1* (APU Report). London: DES.

Foxman, D., Ruddock, G., Joffe, L., Mason, K., Mitchell, P. and Sexton, B. (1985). *Mathematical Development: a Review of Monitoring in Mathematics 1978 to 1982. Part 2* (APU Report). London: DES.

Foxman, D., Ruddock, G., McCallum, I. and Schagen, I. (1991). *APU Mathematics Monitoring (Phase 2)*. London: SEAC.

Foxman, D.D., Badger, M.E., Martini, R.M. and Mitchell, P. (1981). *Mathematical Development: Secondary Survey Report No. 2* (APU Survey). London: HMSO.

Foxman, D.D., Cresswell, M.J. and Badger, M.E. (1981). *Assessment of Performance Unit Mathematical Development: Primary Survey Report No. 2 Supplement (Wales)*. Cardiff: Welsh Office.

Foxman, D.D., Cresswell, M.J. and Badger, M.E. (1981). *Mathematical Development: Primary Survey Report No. 2* (APU Survey). London: HMSO.

Foxman, D.D., Cresswell, M.J., Ward, M., Badger, M.E., Tuson, J.A. and Bloomfield, B.A. (1980). *Mathematical Development: Primary Survey Report No. 1* (APU Survey). London: HMSO.

Foxman, D.D., Martini, R.M. and Mitchell, P. (1982). *Assessment of Performance Unit Mathematical Development: Secondary Survey Report No. 3 Supplement (Wales)*. Cardiff: Welsh Office.

Foxman, D.D., Martini, R.M. and Mitchell, P. (1982). *Mathematical Development: Secondary Survey Report No. 3* (APU Survey). London: HMSO.

Foxman, D.D., Martini, R.M., Tuson, J.A. and Cresswell, M.J. (1980). *Mathematical Development: Secondary Survey Report No. 1* (APU Survey). London: HMSO.

Foxman, D.D., Ruddock, G.J., Badger, M.E. and Martini, R.M. (1982). *Mathematical Development: Primary Survey Report No. 3* (APU Survey). London: HMSO.

Gamble, R., Davey, A., Gott, R. and Welford, G. (1985). *Science at Age 15: a Report on the Findings of the Age 15 APU Science Surveys*. London: APU.

Gipps, C. and Goldstein, H. (1983). *Monitoring Children: an Evaluation of the Assessment of Performance Unit*. London: Heinemann.

Gipps, C. and Goldstein, H. (1983). *Monitoring Children: an Evaluation of the Asssessment on Performance Unit. Supplementary Appendices (6-9)*. London: University of London, Institute of Education.

Gorman, E.P., White, J., Orchard, L. and Tate, A. with Sexton, B. (1979). *Language Performance in Schools: Secondary Survey Report No.1*. London: HMSO.

Gorman, T.P. (1986). *The Framework for the Assessment of Language* (APU Survey). Windsor: NFER-NELSON.

Gorman, T.P. (1987). *Pupils' Attitudes to Readiing at Age 11 and 15* (APU Survey). Windsor: NFER-NELSON.

Gorman, T.P. and Kispal, A. (1987). *The Assessment of Reading: Pupils Aged 11 and 15* (APU Survey). Windsor: NFER-NELSON.

Gorman, T.P., Purves, A.C. and Degenhart, R.E. (Eds) (1988). *The IEA Study of Written Composition 1: the International Writing Tasks and Scoring Scales.* Oxford: Pergamon Press.

Gorman, T.P., White, J. and Brooks, G. (1984). *Language Performance in Schools: 1982 Secondary Survey Report* (APU Survey). London: HMSO.

Gorman, T.P., White, J., Brooks, G. and English, F. (1991). *Language for Learning: a Summary Report on the 1988 APU Language Performance Survey* (Assessment Matters No. 4). London: SEAC.

Gorman, T.P., White, J., Brooks, G., MacLure, M. and Kispal, A. (1988). *Language Performance in Schools: Review of APU Language Monitoring 1979-1983*. London: HMSO.

Gorman, T.P., White, J., Hargreaves, M., MacLure, M. and Tate, A. (1984). *Language Performance in Schools: 1982 Primary Survey Report* (APU Survey). London: DES.

Gorman, T.P., White, J., Orchard, L. and Tate, A. with Sexton, B. (1981). *Language Performance in Schools: Primary Survey Report No. 1* (APU Survey). London: HMSO.

Gorman, T.P., White, J., Orchard, L. and Tate, A. with Sexton, B. (1982). *Language Performance in Schools: Secondary Survey Report No. 1* (APU Survey). London: HMSO.

Gorman, T.P., White, J., Orchard, L. and Tate, A. with Sexton, B. (1982). *Language Performance in Schools: Primary Survey Report No. 2* (APU Survey). London: HMSO.

Gorman, T.P., White, J., Orchard, L. and Tate, A. with Sexton, B. (1983). *Language Performance in Schools: Secondary Survey Report No. 2* (APU Survey). London: HMSO.

Gott, R. (1984). *Electricity at Age 15: a Report on the Performance of Pupils of Age 15 on Questions in Electricity* (Science Report for Teachers: 7). London: DES.

Gott, R., Davey, A., Gamble, R., Head, J., Khaligh, N., Murphy, P., Orgee, T., Schofield, B. and Welford, G. (2008). *Science in Schools Ages 13 and 15: Report No 3*. London: DES.

Gott, R. and Murphy, P. (2008). *Assessing Investigations at Ages 13 and 15: a Report for Teachers on the Planning and Performance of Investigations by Pupils of Ages 13 and 15*. London: DES.

Harlen, W. (1986). *Planning Scientific Investigations at Age 11* (Science Report for Teachers: 8). London: DES.

Harlen, W., Black, P. and Johnson, S. (1981). *Science in Schools Age 11: Report No 1*. London: HMSO.

Harlen, W., Black, P., Johnson, S. and Palacio, D. (1983). *Science in Schools Age 11: Report No 2*. London: HMSO.

Harlen, W., Black, P., Johnson, S., Palacio, D. and Russell, T. (1984). *Science in Schools Age 11: Report No 3*. London: HMSO.

Harlen, W., Johnson, S. and Palacio, D. (1983). *Science at Age 11* (Science Report for Teachers: 1). London: DES.

Harlen, W., Palacio, D. and Russell, T. (1984). *Science Assessment Framework Age 11* (Science Report for Teachers: 4). London: HMSO.

Joffe, L. (1982). *Practical Testing in Mathematics at Age 15*. London: DES.

Joffe, L. (1985). *Practical Testing in Mathematics at Age 11*. London: DES.

Joffe, L. and Foxman, D. (1988). *Attitudes and Gender Differences: Mathematics at Age 11 and 15*. Windsor: NFER-NELSON.

Joffe, L. and Foxman, D. (1989). *Communicating Mathematical Ideas: a Practical Interactive Approach at Ages 11 and 15*. London: HMSO.

Johnson, S. (1989). *National Assessment. The APU Science Approach*. London: HMSO.

Johnson, S. and Murphy, P. (1986). *Girls and Physics: Reflections on APU Survey Findings* (APU Occasional Paper No 4). London: DES.

Kelly, A.V., Kimbell, R.A., Patterson, V.J., Saxton, J. and Stables, K. (1987). *Design and Technological Activity: a Framework for Assessment*. London: HMSO.

Keys, W. and Foxman, D. (1989). *A World of Differences: a United Kingdom Perspective on an International Assessment of Mathematics and Science*. Slough: NFER.

Khaligh, N., Johnson, S., Murphy, P., Orgee, T. and Schofield, B. (1986). *Science in Schools Age 13: Report No 4*. London: DES.

Kimbell, R., Stables, K., Wheeler, T., Wosniak, A. and Kelly, V. (1991). *The Assessment of Performance in Design and Technology: the Final Report of the APU Design and Technology Project 2 1985-1991*. London: SEAC.

Lee, B. (1987). *Foreign Languages: Writing. Implications of the APU Surveys for Teaching and Learning*. Windsor: NFER-NELSON.

MacLure, M. and Hargreaves, M. (1986). *Speaking and Listening: Assessment at Age 11* (APU Survey). Windsor: NFER-NELSON.

Mason, K. and Ruddock, G. (1986). *Decimals: Assessment at Age 11 and 15*. Windsor: NFER-NELSON.

Mason, K. and Tooley, J. with Foxman, D. (1992). *Moving Forward in Maths: a Diagnostic Teaching Approach. Number*. Windsor: NFER-NELSON.

Murphy, P. and Gott, R. (1984). *Science Assessment Framework Age 13 & 15* (Science Report for Teachers: 2). London: DES.

Murphy, P. and Schofield, B. (1984). *Science at Age 13* (Science Report for Teachers: 3). London: DES.

National Foundation for Educational Research (1983). *Foreign Language Provision: Survey of Schools, Autumn 1982*. London: DES.

Rosen, H. (1982). *The Language Monitors: a Critique of the APU's Primary Survey Report Language Performance in Schools* (Bedford Way Papers 11). London: University of London, Institute of Education.

Ruddock, G. (1987). *The Cockcroft Foundation List: APU Results*. Windsor: NFER-NELSON.

Russell, A., Black, P., Bell, J. and Daniels, F. (1991). *Observation in School Science: a Booklet for Teachers* (Assessment Matters No. 8). London: SEAC.

Russell, T. (Ed) (1988). *Science at Age 11: a Review of APU Survey Findings.* London: HMSO.

Schofield, B. (Ed) (1989). *Science at Age 13: a Review of APU Survey Findings 1980-84.* London: HMSO.

Schofield, B., Black, P., Head, J. and Murphy, P. (1984). *Science in Schools Age 13: Report No 2*. London: HMSO.

Schofield, B., Murphy, P., Johnson, S. and Black, P. (1982). *Science in Schools Age 13: Report No 1*. London: HMSO.

Sexton, B. (1981). *A Technical Supplement on the Analysis of APU Monitoring Language*. Slough: NFER.

Stow, M. with Foxman, D. (1989). *Mathematics Coordination: a Study of Practice in Primary and Middle Schools*. Windsor: NFER-NELSON.

Strang, J. (1990). *Measurement in School Science* (Assessment Matters No. 2). London: SEAC.

Strang, J., Daniels, S. and Bell, J. (1991). *Planning and Carrying Out Investigations: a Booklet for Teachers* (Assessment Matters No. 6). London: SEAC.

Taylor, R.M. and Swatton, P. (1990). *Graph Work in School Science: a Booklet for Teachers* (Assessment Matters No. 1). London: SEAC.

Thornton, G. (1987). *APU Language Testing 1979-1983. An Independent Appraisal of the Findings*. London: HMSO.

Welford, G., Bell, J., Davey, A., Gamble, R. and Gott, R. (1986). *Science in Schools Age 15: Report No 4*. London: DES.

Welford, G., Harlen, W. and Schofield, B. (1985). *Practical Testing at Ages 11, 13 and 15: a Report on the Testing of Practical Skills in Science at Three Ages as Undertaken by the Science Team of the APU* (Science Report for Teachers: 6). London: DES.

Wells, I.F. (1982). *School Mathematics in Perspective: an Interpretation by the Secondary School Mathematics Community of Northern Ireland of the First APU Report on the Performance of Fifteen-Year olds*. Belfast: Northern Ireland Council for Educational Research.

White, J. (1986). *The Assessment of Writing, Pupils Aged 11 and 15* (APU Survey). Windsor: NFER-NELSON.

White, J. (1987). *Pupils' Attitudes to Writing at Age 11 and 15* (APU Survey). Windsor: NFER-NELSON.

White, J. and Welford, G. (1988). *The Language of Science* (Science Report for Teachers: 11). London: DES.