

# **Considerations for future assessment of Key Stage 2 science**

**National Foundation for Educational Research**

**November 2009**





# Contents

<b>1</b>	<b>Executive Summary .....</b>	<b>2</b>
1.1	Case Studies .....	3
1.2	Discussion .....	8
1.3	Recommendations .....	10
<b>2</b>	<b>Introduction .....</b>	<b>15</b>
	National Monitoring Case Studies .....	17
2.1	Case Study 1: National Education Monitoring Project (NEMP) in New Zealand.....	17
2.2	Case Study 2: The National Assessment of Educational Progress (NAEP) in the USA	18
2.3	Case Study 3: Scottish Survey of Achievement (SSA).....	21
2.4	Case Study 4: The Assessment of Performance Unit (APU) in England.....	21
	Teacher Assessment Case Studies .....	25
2.5	Case Study 5: Scientific Minds – A guide to assessing attainment target one.....	25
2.6	Case Study 6: Year 4 science optional tasks .....	26
2.7	Case Study 7: Assessing Progress in Science .....	28
<b>3</b>	<b>Discussion.....</b>	<b>31</b>
3.1	National monitoring .....	31
3.2	Teacher Assessment.....	37
<b>4</b>	<b>Recommendations .....</b>	<b>41</b>
4.1	National monitoring .....	41
4.2	Teacher Assessment.....	43
<b>5.</b>	<b>References .....</b>	<b>45</b>
	<b>Appendix 1: NFER Credentials.....</b>	<b>47</b>

# 1 Executive Summary

In May 2009, *The Report of the Expert Group on Assessment* was published (Bevan *et al.*, 2009) and recommended an end of whole cohort testing in science at the end of Key Stage 2. Ed Balls MP, Secretary of State for Education, accepted the changes recommended by the Expert Group and proposed in addition that schools in England should have a system of national sampling at Key Stage 2. In response, a monitoring test will be run in 2010 and 2011 using traditional National Curriculum test papers on a sample of the cohort but in 2012 a new, specifically designed testing system at Key Stage 2 will be introduced.

The Expert Group also concluded that teaching and assessment of science can be improved and its importance in the primary curriculum reinforced by replacing externally marked tests by high quality teacher assessments for reporting on performance of individual pupils. The aim of these high quality teacher assessments would be to recognise whether pupils have a firm grip on the practical nature of science, and the skills to develop and apply scientific understanding. Teacher assessment is more likely to be able to assess the practical elements of the science curriculum.

This paper by the National Foundation for Educational Research (NFER) aims to highlight the key issues that need to be addressed when considering the introduction of a national monitoring system in science and high quality teacher assessment tasks. The NFER has conducted a review of practical investigative tasks in science from other countries and from past experience in England. This paper details the key features of these systems and the implications for the introduction of such an approach at Key Stage 2. The paper provides brief case studies of:

- National Education Monitoring Project (NEMP) in New Zealand;
- 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 Assessment of Performance Unit (APU), the national monitoring system used in England up until 1989;
- *Scientific Minds – A Guide to Assessing Attainment Target One*, (Jones and Griffin 2005), published by nferNelson;
- Year 4 science optional tasks (QCA 1997);
- *Assessing Progress in Science* (Russell and McGuigan, 2003), published by QCA.

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 considerations arising relevant to this development in science. The main purpose of the case studies is to inform and provide evidence to support the discussion section later in this paper. The case studies have been divided into two groups: examples of national monitoring systems with a practical component, and examples of

tasks used as part of teacher assessment. Key lessons from each of the case studies are given in Section 1.1 below.

This report aims to highlight lessons to be learnt from existing, established assessments about the assessment of practical, investigative science. A number of current developments in science assessment, such as the introduction of Assessing Pupil Progress (APP) in England, a wider use of cross-curricular study, and of the use of ICT to support teaching, learning and assessment in science, have not been discussed.

As part of this research, some questions were included in NFER's Teacher Voice Omnibus survey in June 2009, asking primary teachers about the impact of no longer using the National Curriculum tests to assess Key Stage 2 science and how they would like to see science assessed in the future.

## **1.1 Case Studies**

A number of examples of investigative science tasks have been reviewed as part of the development of this report. Each example has been described in a case study in the main body of the report. The case studies each raise a number of key lessons that ought to be considered as we introduce similar assessments in England. A brief overview of the case study assessments and the lessons learnt from each are provided below, divided into those for national monitoring surveys and those for practical teacher assessment tasks.

### **National Monitoring**

#### **1.1.1 NEMP Key Lessons (New Zealand)**

National monitoring in New Zealand was introduced in general education settings in 1995 and in Māori Medium settings in 1999. It aims to provide dependable information about the achievements of New Zealand pupils, using approaches tailored to New Zealand's school system and curriculum. The Educational Assessment Research Unit of the University of Otago, contracted by the Ministry of Education, runs the project. Below are some key lessons based on the assessment of science in NEMP.

- 1 The survey includes tasks which enable practical skills to be assessed. This is particularly valuable in a practical subject such as science.
- 2 The tasks are short, manageable and engaging for pupils. Teacher-administrators find them straightforward to mark, simply recording which of a number of possible responses a pupil/group gives.
- 3 Teachers who train as administrators highly value the professional development they receive.
- 4 Researchers watching videos of the assessments have noticed that the teacher-administrators vary in the amount of assistance they give pupils.
- 5 The training of administrators and production of materials involves significant expenditure.

- 6 Pupils seem to be benefitting from the project less than the teacher administrators: an increasing number of pupils say in the surveys that they ‘never’ do practical investigations in science.
- 7 Some problems have arisen with the random sampling of pupils: selected pupils who are recent immigrants or disabled are often replaced.

### **1.1.2 NAEP Key Lessons (USA)**

The National Assessment of Educational Progress surveys in the USA are carried out with pupils in Grade 4 (aged 9-10) and also with pupils in Grades 8 and 12. It is the assessments for Grade 4 pupils that are considered in this case study. Information on pupil performance is reported at national and state level and also for some large urban areas. Results are broken down by pupil groups, including gender, SES and ethnicity. Trends over time are reported.

For considerations relating to NAEP as a national monitoring model, please see the NFER submission to the Expert Group (NFER, 2009). Below are some considerations for using hands-on tasks and Interactive Computer Tasks (ICTs) in science assessment.

- 1 The hands-on tasks used in the survey in addition to paper and pencil tests enable a broader testing of scientific inquiry and skills which are closer to classroom experiences and reflect the importance of ‘doing’ science.
- 2 Challenges of this method of assessment are the time to carry out the assessments, the difficulty of creating tasks, equipment costs and scoring costs.
- 3 The most recent survey included a number of Interactive Computer Tasks (ICT) which were hands-on and engaging for pupils.
- 4 ICTs can be used to do things not possible in other formats, such as show processes in slow motion and assess skills in finding information; they are potentially cost effective; and can be used to simulate experiments that require a large number of readings, such as rolling a ball down different height slopes with different surfaces.
- 5 Concerns of using computer-based assessments include the availability of computers, limited research into computer-based assessment and initial development costs.

### **1.1.3 SSA Key Lessons (Scotland)**

The Scottish Survey of Achievement (SSA) was introduced in 2005 to replace the Assessment of Achievement Programme (AAP) which had been running since 1983. The SSA is run by the Scottish Government, in partnership with the Scottish Qualifications Authority (SQA) and Learning and Teaching Scotland (LTS).

For considerations relating to SSA as a national monitoring model, please see the NFER submission to the Expert Group (NFER, 2009). Below are some considerations for using practical science tasks as part of a national monitoring assessment.

- 1 In the recent past the core SSA has used paper and pencil tests to assess across the curriculum in particular subject areas, including science. In 2007 the tests were designed to

assess pupils' science knowledge and understanding. The paper and pencil testing was supplemented by very much smaller scale practical assessments, whose results were intended to be indicative of pupils' achievements, providing additional information on specific areas, and developing assessment expertise in the teaching profession.

- 2 Teachers who served as field officers felt strongly that they benefited from CPD.
- 3 Rating schemes were used by field officers for levelling pupils for some types of practical assessment; with training provided to ensure as far as possible that the field officers interpreted the schemes consistently.
- 4 For the informal assessment of science investigation skills that featured in the 2007 survey, field officers did not actually see the pupils doing the practical investigations; but discussed these with the pupils when they visited the schools at a later date.
- 5 Teacher assessments are also collected for the group of pupils being tested. Disparities are seen between the test results and the teacher judgements, especially in science (see Johnson and Munro, 2008).
- 6 Teachers are involved in the process, including for marking some aspects of the work, for assessing certain skills, and for providing teacher assessments based on class work. This capacity building within the teaching community can be seen as a very useful additional benefit of introducing national monitoring surveys.

#### **1.1.4 APU Key Lessons (England and Wales)**

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 Department of Education and Science (DES) to run the project. 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.

For considerations relating to APU as a national monitoring model, please see the NFER submission to the Expert Group (NFER, 2009). Below are some considerations for using practical science tasks as part of a national monitoring assessment.

- 1 The small scale, one to one assessment of practical investigation skills in science, while ground-breaking, was time consuming and expensive. In addition, training for assessors was required. The 'circus' model was less time consuming.
- 2 Pupils were assessed individually for the 'performing investigations' practical component.
- 3 The administration of the tests with only seven pupils in each school made them logistically difficult to manage.

#### **Teacher Assessments**

The following section provides an overview of examples of teacher assessment tasks and the key lessons learnt from them.

### **1.1.5 Scientific Minds Key Lessons (nferNelson)**

*Scientific Minds* is a publicly available teacher resource produced by nferNelson in 2005, developed by NFER. Materials are available for key stage 1, key stage 2 and key stage 3. The case study here will focus on the key stage 2 materials (Jones and Griffin, 2005).

The Key Stage 2 book describes techniques that teachers can use to assess pupils' performance in Attainment Target 1 (AT1). The book contains four different activities. Each activity consists of a list of materials needed, a description of the investigation, links to the National Curriculum Programme of Study, and provides guidance on how to assess pupils' performance and how to level pupil work. The aim of the book is to give teachers ideas about how assessment of AT1 can be carried out, and to help teachers use the results of assessment meaningfully.

- 1 Each assessment lists the equipment required for each task.
- 2 A summary is provided for teachers with regards to length of activity, the purpose of the activity, which parts of the Programme of Study are being assessed etc.
- 3 Examples of 'pupil speak' answers are provided which make it clear what teachers should be looking for.
- 4 Pre-prepared worksheets which are specifically designed to assess particular areas of the curriculum are also included.
- 5 The tasks are challenging and the ways in which pupils present their work innovative and interesting.
- 6 Practical activities were trialled in order to ensure they work and any equipment needed is accessible to primary teachers.

### **1.1.6 Year 4 science optional tasks Key Lessons (QCA)**

This series of assessments were designed to assess pupils halfway through Key Stage 2 (at the end of Year 4). They were written to assess English, mathematics and science. The focus of the case study is the science materials. Five units were developed and published in late 1997 to assess pupils working at Levels 2-4 across the Key Stage 1 and Key Stage 2 Programmes of Study. Each unit has two separate activities. These materials were developed by NFER during 1997 and published by QCA in late 1997.

- 1 There is information contained within the Teacher's Guide to support the administration of the materials and to help teachers interpret the results with respect to level. Exemplar materials for pupils performing at different levels are included.
- 2 The flexibility of the assessment is a strength, so that teachers can judge when pupils are likely to perform at their best.
- 3 Teachers were able to pick and choose which activities they wanted to use.
- 4 Each activity clearly sets out what the teacher needed to do, what prompts to give where necessary, what answers were expected and some example responses. Work sheets are given for some of the activities.
- 5 The materials only partially assessed the curriculum, so teachers could not rely on these as the sole means of assessment.

- 6 Teachers have the option to ask for clarification of responses given by pupils.
- 7 The variety of activities, including practical investigations, is a strength of this set of materials.

### **1.1.7 Assessing Progress in Science Key Lessons (QCA)**

This set of materials is designed to encourage teachers to incorporate assessment into the teaching and learning cycle. If teachers are able to gather information on what has been learnt, this information can be used to adapt future lessons. The assessment aims to provide diagnostic information to teachers and pupils.

The materials cover a broad sweep of the curriculum at Key Stage 2, but they do not assess the entire curriculum. There are eight units in Key Stage 2 assessing elements of Sc1 - Sc4. The assessment activities themselves each have a teaching and learning sequence followed by a number of assessment activities. Notes are provided on reviewing and interpreting the evidence presented by the pupils.

The materials were produced in 2003 by Qualifications and Curriculum Authority (QCA) and were written by researchers at the Centre for Research into Primary Science and Technology (CRIPSAT).

- 1 The teacher's guide is quite theoretical; this may be helpful if the aim is to increase teachers' knowledge and understanding of the different types of assessment.
- 2 Providing ideas sheets may help pupils to pin down any ideas they have and the teacher has a record of the original thoughts if they want to ask the pupil to clarify any of their ideas.
- 3 Examples of answers that pupils tend to give are provided as are types of responses given by pupils working at different levels.
- 4 Along with methods of interpreting answers with respect to the National Curriculum, there is some advice on feeding back to pupils, although this guidance tends to be very general and it is difficult to know whether teachers would find this type of advice useful.
- 5 The activities are not explicitly linked to the Programme of Study. Where references are given, it is to the Schemes of Work. When describing performance at a level, there are some references to the level descriptors.
- 6 Examples of pupil responses in the level charts are not given in 'pupil speak'. Giving examples in 'pupil speak' may help teachers interpret responses more reliably.
- 7 The activities are manageable for teacher assessment in terms of time and resources.
- 8 There is an appropriate range of activities across the science curriculum, including practical and written components.

## **1.2 Discussion**

The discussion section gives the implications of the case studies for the proposed developments. It first focuses on national monitoring assessments in science and then considers teacher assessment tasks in science.

### **1.2.1 Purpose**

The first critical decision to be made with respect to national monitoring in science at Key Stage 2 must relate to 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. If monitoring systems for Key Stages 2 and 3 are designed taking the other key stage into account, it may be possible to monitor development in science between the two key stages. 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.

As there is some concern across the science community that there is a decreased emphasis being placed on science in the future years (because of the fact that there is no longer statutory assessment in science and its loss of ‘core subject’ status in the revised primary curriculum), monitoring is seen as important so that any downward changes in standards can be highlighted early. It will also serve to monitor any increases in standards because of schools’ increased flexibility with respect to the curriculum and classroom practice.

The purpose of teacher assessment is likely to be to provide information to feed into teaching and learning in the classroom. It is also likely to be used for reporting to parents. If the information is to be reported to other stakeholders, the process of assessment may need to be moderated. Teacher assessments should not be used for assessments used as part of an accountability system, either for teacher, school or national accountability.

### **1.2.2 Structure**

The structure of an assessment relates to how pupils would take it; as a pencil and paper test, as a computer-based assessment, as a practical test or a combination of types of assessment. This will be determined by which areas of the curriculum are deemed a priority to assess and how often. A more complex moderation and marker training system is likely to be needed for practical and open response assessments than for objective, computer marked tests (more complex computer-based items may well require the same systems as those not done on computer).

Manageability, methods of recording the data and impact on consistency of the resulting data will also influence the structure of the tests. An assessment that is simple to administer is more likely to be administered consistently across a range of test sites. This does, however, need to be balanced with collecting the appropriate information to meet the purposes of the assessment.

The structure of the assessment will be affected by the extent to which the entire curriculum is being assessed. Assessing most of the curriculum in each assessment cycle will produce either longer tests or a number of test versions. Increasing the number of test versions decreases the burden on any one school, but more schools will be needed to obtain the sample required. Having longer tests will increase the burden on individual pupils.

In teacher assessment, the tasks need to be realistic and achievable, assess areas across the curriculum in a variety of ways, provide information to help teachers in making judgements and possibly most importantly, provide a model teachers can use to develop their own assessments.

### **1.2.3 Sampling**

In determining the sample, there is a tension between minimising the total number of schools/pupils required to be assessed each time, and having sufficiently precise data to report reliably on any sub-groups of the population being monitored and/or on sub-sections of the curriculum. If data were to be reported to individual LAs, for example, there would need to be sufficient numbers assessed in each LA to produce this information; many more pupils would be needed than would be required if only national data were reported.

A monitoring assessment is likely to be low stakes at the time of administration, especially if schools do not receive any feedback on their pupils' performance. Care would need to be taken if making comparisons with the National Curriculum assessment data for this reason.

Practical assessments are more time consuming due to the number of pupils that can be legitimately assessed at any one time. Training and moderating for practical assessments add to the time requirement. There are also logistical and financial considerations related to practical assessments.

Sampling is not generally needed for teacher assessments, although it may be for their moderation.

### **1.2.4 Administration and Manageability**

Administration and manageability are closely linked to the structure of the assessments. Pencil and paper or computer-based assessments are likely to be relatively simple to administer when compared with a practical assessment (assuming the technical infrastructure is in place for computer-based tests). Greater numbers of test versions or components could increase errors in administration and increase the amount of data that cannot be used.

As practical assessments are more difficult to observe and make consistent judgements across a number of administrations, it would be preferable to use trained administrators. Using administrators means that a smaller group of people need to be trained, the observations are more consistent and it decreases the burden on schools. However, the likely increased consistency of the results must be compared with the professional development benefits of using class teachers.

It may be possible to achieve both aspects by using class teachers, with a limited leave of absence from the classroom, acting as field officers.

Teacher assessment tasks need to be trialled to ensure that they are workable in the classroom under a range of classroom situations. Providing lists of equipment and simple, clear instructions will also help the manageability of the assessments. It is essential that information is provided in terms of interpreting pupil responses, so that teachers are able to produce useful information for teaching and learning.

### **1.2.5 Marking and Reporting**

In a monitoring assessment, marking requirements will obviously depend on the type of assessment. For scripts, marking centres are useful for marking a large number of scripts quickly. They also have the advantage that the quality of marking can be monitored during the marking process. However, as marking centres would only be in a limited number of areas, markers would have to be local to those sites. This makes it more difficult to train markers from a range of schools to improve skills in assessment.

In a system that uses computer-based assessment, marking could be largely automated, with perhaps professional marking for a small proportion of questions.

A practical component to a science monitoring assessment presents the most difficulties in terms of training administrators/observers, recording observations and maintaining consistency across a large number of schools.

In terms of teacher assessment, tasks should be structured to allow marking and levelling to be conducted by teachers as consistently as possible. A number of the case studies use ‘pupil speak’ responses to aid in this.

## **1.3 Recommendations**

### **National monitoring**

#### **1.3.1 Purposes**

Given the importance of science within the context of the national STEM (Science, Technology, Engineering and Mathematics) agenda, monitoring will act as a check that the amount and quality of science teaching in primary schools do not suffer now that the tests are no longer statutory. Monitoring will ensure that standards are maintained over time and that the quality of science teaching does not suffer because of a perceived loss of emphasis on science as a subject. It may be useful to expand on this further and consider questions such as whether standards need to be linked to those from previous Key Stage 2 tests, in which case there will need to be means of linking the two forms of assessment, and there will need to be similarities between the assessments. However, the usefulness of such linking could be questioned given the different

nature and purposes of the two sets of assessments. One key aspect that will need to be taken into account if any linking is attempted, is the motivation of the pupils taking each assessment. If the purpose is to assess a wider range of skills in science then it may be more appropriate to start measuring standards from this point in time and to have a different approach to the assessment. In reality it may be more appropriate to aim for a combination of the two. So there are some similar items that can be used for linking, with new items, such as more open-ended tasks, added. As it is likely that the key stage tests that were being developed for statutory testing will be used for national monitoring in 2010 and 2011, it may be possible to link the monitoring tests to the 2009 statutory test.

- 1 A significant part of the science curriculum is the practical/investigative element and the criticisms of the current tests largely focused on the absence of this element. A national sample monitoring system which includes a practical element would be a better assessment of the full science domain. It also may encourage more practical investigative work in science teaching.
- 2 It is difficult to know how monitoring will impact on teaching and learning in classrooms; the monitoring tests will be lower stakes than the National Curriculum tests. If schools do not receive results, monitoring may have little effect on school behaviour. However, as the National Curriculum tests were high stakes and have been said to have adversely impacted teaching and learning, it is argued here that the monitoring tests should not attempt to influence what goes on in the classroom through the tests.

### **1.3.2 Structure**

- 3 When designing the structure of the assessment, methods to ensure the consistency of administration, marking and interpreting the data need to be built into the system. This may include, for example, using administrators in practical assessments.
- 4 Including a practical component will improve the validity of the survey as an assessment of science and will ensure that it meets the expectations of a wider group of stakeholders providing the sample numbers are fit for purpose. We therefore believe this will be an important component of the new assessment, although it will increase the costs of running the monitoring survey.
- 5 Assessing most of the curriculum in each assessment cycle will provide a snapshot of performance in science. Assessing a focused part of the curriculum would mean that several cycles would have to contribute and a picture of performance at any one time could not be easily ascertained.
- 6 It is recommended that the stakes of the tests be kept low as far as possible, but allowance be 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.
- 7 If monitoring surveys across Key Stages 2 and 3 are designed in conjunction, it may be possible to look at development across the two key stages.

### 1.3.3 Sampling

- 8 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 assessments.
- 9 It was recommended in the NFER report on national monitoring in general (NFER, 2009) 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. In the case of practical assessments it is likely that a subset of the main sample will be used. However, if the practical tasks are to form a high profile aspect of the system then it is important that the sample be sufficiently large to provide statistically reliable results.

### 1.3.4 Administration and Manageability

- 10 The burden on schools taking part in any national survey should not be onerous and ideally, schools should have something in return for participating as recognition that they have done so. This does not necessarily have to be financial; rewards may be offered in terms of professional development for example. Any demands on school staff as a result of participation in the surveys should be kept to a minimum and should be designed to fit within school organisation as far as possible. Practical tasks do tend to be more burdensome both in the resources and organisation required to administer them, but they would be seen by the science community as a better measure of science than the National Curriculum assessments. For this reason, a practical element should be included in a national monitoring assessment. Teachers could be trained as administrators which would then impact positively on classroom practice.
- 11 It is recommended that the use of technology be considered for both the administration of the tests and the marking. This could allow the use of more complex data in science, or simulations or video in slow motion. It may also be possible to collect the background data on the pupils and attitudes to learning in the form of an online survey.
- 12 Paper and pencil or computer-based assessments could possibly be administered by school staff. Trained administrators should be used for any practical assessments, for consistency across administration centres and in interpreting observations.
- 13 Some of the case studies (e.g. NEMP) involve assessment of group work. This has some advantages if group working skills are being assessed, although it is difficult to assess any one particular pupil in a group working environment. NEMP uses a mixture of group working and individual working administrations to get around these issues. If group working was identified as an important part of the science curriculum, then a mixture of administration types may be more appropriate.

### **1.3.5 Marking and Reporting**

- 14 Marking pencil and paper assessments and practical assessments requires trained markers and observers and/or moderation of the marking to ensure that there is consistency across the cohort. Any practical administrations carried out by observers must consider how much any one observer can reliably do at one time; observations must either be simple and easily recorded or more complex and recorded in some way to allow the observer to replay it (e.g. video or audio recording). The number of pupils being observed will also impact on how the data is collected.
- 15 As with many aspects of this assessment, a clear definition of the purpose and what will be reported will determine the number of pupils that will need to be practically assessed. If the numbers of pupils are too small in any subset, it may not be possible to report valid information.
- 16 It is recommended that the survey design includes ways of linking the results to results from the TIMSS survey at grade 4 as this is the closest to Key Stage 2, to allow international comparisons to be made. The linking should take the form of common tasks or tests.

#### Teacher Assessments

### **1.3.6 Purposes**

- 1 High quality teacher assessment tasks should provide a variety of activities based around practical investigative science which can be used flexibly to support teaching and learning. Teacher assessment tasks should not aim to cover the whole curriculum as teachers should not be able to rely on them as the sole means of assessment.

### **1.3.7 Structure**

- 2 Use a variety of activities incorporated in teaching and learning cycles to assess different parts of the curriculum.
- 3 Provide clear guidance on how to carry out tasks and assess pupils; including concise instructions, length of the task, any prompts needed, worksheets, marking guidance and examples of pupil responses.

### **1.3.8 Administration and manageability**

- 4 Tasks must have been trialled in realistic situations.
- 5 Allow flexibility for teachers to make adjustments to suit their own context.
- 6 Include ready-to-use worksheets but provide opportunities for other methods of information collection which teachers can use to level a pupil.

### **1.3.9 Marking and reporting**

- 7 Provide information in the form of descriptors of performance at each level and pupil exemplars to help teachers interpret results with respect to level in each task. These examples can be provided by pre-testing or trialling the tasks prior to publication.
- 8 Provide guidance on giving feedback to pupils and identifying the next steps. Feedback could include an assessment of what the pupil has answered well, identifying areas of weakness and next learning steps.

### **1.3.10 Other considerations**

- 9 Provide examples of tasks and how to assess them so teachers are able to develop their own assessment materials.

## 2 Introduction

In May 2009, *The Report of the Expert Group on Assessment* was published (Bevan *et al.*, 2009). It recommended an end of whole cohort testing in science at the end of Key Stage 2. Ed Balls MP, Secretary of State for Education, accepted the changes recommended by the Expert Group and proposed in addition that schools in England should have a system of national sampling at Key Stage 2. In response a monitoring test will be run in 2010 and 2011 using traditional National Curriculum test papers on a sample of the cohort but in 2012 a new, specifically designed testing system at Key Stage 2, will be introduced. The national monitoring surveys will be valuable for tracking changes to standards over time, which may arise as science is no longer the subject of high stakes testing, and for maintaining the importance of science within the primary curriculum.

The Expert Group concluded that teaching and assessment of science can be improved and its importance in the primary curriculum reinforced by replacing externally marked tests by high quality teacher assessments for reporting on the performance of individual pupils. The aim of these high quality teacher assessments would be to recognise whether pupils have a firm grip of the practical nature of science, and the skills to develop and apply scientific understanding. The Expert Group states that '*Externally set and marked tests cannot readily assess investigative skills and the ability to design and carry out an experiment and understand its results*' (Bevan *et al.*, 2009, p. 25); this is a view widely held throughout the science and primary teaching community. Teacher assessment is more likely to be able to assess the practical elements of the science curriculum. The group also recognised that teacher assessment '*cannot be used as part of the accountability framework at Key Stage 2*' (p. 25).

The Royal Society, SCORE, the Science Learning Centres, the Association of Science Educators, Wellcome and the Gatsby Charitable Foundation have all agreed to provide advice on the nature of assessment in Key Stage 2 science with the aim of strengthening teaching and learning in primary science.

NFER (the National Foundation for Educational Research) has drafted this paper on lessons learnt from its experience of assessment development over a number of years to contribute to this debate. NFER's experience includes being the test development agency for the Key Stage 2 National Curriculum tests from 2000 - 2009, administering the international surveys (TIMSS, PIRLS and PISA), and developing teacher assessments. Our recommendations are also informed by reviewing assessment of practical aspects of science in other countries. Full details of NFER's credentials are included as Appendix 1.

This paper by NFER aims to highlight the key issues that need to be addressed when considering the implementation of a national monitoring system in science and high quality teacher assessment tasks with a particular focus on practical, investigative science. In order to demonstrate the different approaches taken with regards to national monitoring and teacher assessments a number of case studies have been included. We have also put forward our

recommendations for the nature of practical science tasks as a part of national monitoring and teacher assessments based on the findings from the case studies. At this stage of the discussion the recommendations must be viewed as possible solutions, as some major decisions need to be made before these can be finalised. The development process for any new system is likely to be one of iteration, where initial decisions impact on future discussions and choices.

The case studies referring to national monitoring are the National Education Monitoring Project (NEMP) in New Zealand, the National Assessment of Educational Progress (NAEP) in the USA, the Scottish Survey of Achievement (SSA) and the Assessment of Performance Unit (APU) in England.

The case studies referring to teacher assessment tasks are Scientific Minds – A guide to assessing attainment target one (UK), Science optional tasks for Key Stage 2 and Key Stage 3 (England and Wales), Year 4 science optional tasks (England and Wales) and Assessing progress in science (England).

The NFER Teacher Voice Omnibus survey (June 2009) asked over 1000 teachers about their preferences for assessing science in the future and the impact of no longer having the end of key stage tests. Feedback from this survey is also included.

This report aims to highlight lessons to be learnt from existing, established assessments. A number of current developments in science assessment, such as the introduction of Assessing Pupil Progress (APP) in England, a wider use of cross-curricular study, and of the use of ICT to support teaching, learning and assessment in science, have not been discussed. NFER is also currently conducting research into the implementation of APP at Key Stage 3. Further information on this project can be found at [www.nfer.ac.uk/nfer/research/projects/pupils-progress-science/pupils-progress-science\\_home.cfm](http://www.nfer.ac.uk/nfer/research/projects/pupils-progress-science/pupils-progress-science_home.cfm).

## ***National Monitoring Case Studies***

### **2.1 Case Study 1: National Education Monitoring Project (NEMP) in New Zealand**

#### **2.1.1 Overview**

National monitoring in New Zealand was introduced in general education settings in 1995 and in Māori Medium settings in 1999. It aims to provide dependable information about the achievements of New Zealand pupils, using approaches tailored to New Zealand's school system and curriculum. It was recommended by at least four national working parties or committees of enquiry operating between 1962 and 1990. The Educational Assessment Research Unit of the University of Otago, contracted by the Ministry of Education, runs the project.

#### **2.1.2 Purpose of the assessment**

The purpose of the assessment is to get a broad picture of the achievements of New Zealand school pupils. This is done by using representative samples of pupils at successive points in time so that:

- trends in educational performance can be identified and reported;
- good information is available to assist policy makers, curriculum specialists and educators with their planning;
- the public are provided with information about trends in educational achievement (knowledge and skills), attitudes and motivation of New Zealand pupils.

NEMP does not produce information about individual pupils, teachers or schools. Common tasks are used in consecutive surveys to report on performance change or stability over time.

#### **2.1.3 Assessment aspects**

Year 4 (age 8-9: half-way through primary school) and Year 8 (age 12-13: end of primary school) pupils in about 260 schools take part in NEMP. (In Māori Medium settings, the focus is at Year 8 level only). Schools are grouped according to area, state-funded or private etc. and then randomly selected. Three thousand pupils in all are randomly selected from these schools.

Each learning area is assessed every four years: science has been assessed in 1995, 1999, 2003 and 2007.

National monitoring includes the use of assessment tasks. There are also paper and pencil tests. This case study will focus on the assessment tasks.

In order to ensure professional and community interests are taken into account, schools, teachers and pupils are involved with task development and trialling. The tasks are designed to be meaningful and enjoyable for the pupils. They include a wide range of activities, from those which the majority of Year 4 pupils are likely to have mastered to those which show the highest achievements of the most capable Year 8 pupils. They take full account of differences of language, culture, gender, ability and disability in their design and administration.

About 100 teachers each year are seconded from schools for a week of training followed by five weeks administering the tasks in the selected schools. Task instructions are given to pupils orally from a script by the teacher-administrators, through video presentations (often used to set the context), on laptop computers, or in writing (Gilmore, 1999).

For the science assessments in NEMP, pupils work on tasks with the support of a trained teacher-administrator, in three different ways:

1. *One-to-one*: One pupil working with a teacher – these can be practical investigations or observations, answering questions about a picture or a video clip or playing a game.
2. *Group*: Four pupils working cooperatively – these are mostly practical investigations e.g. planning an investigation, carrying out a fair test, making observations. The group is assessed as a whole.
3. *Stations*: Four pupils working independently around a series of ‘hands-on’ activities e.g. classifying using stickers, choosing correct words from a box, ordering, and answering questions about pictures.

Each pupil works for about three to four hours spread over a period of five days. Some tasks are video-taped to enable detailed analysis later on.

The administrators mark pupil responses, which are presented orally, by demonstration, in writing, in computer files, or through submission of other physical products. For each task the final report compares how Year 4 and Year 8 pupils performed, and records how pupil performance compares to the previous years in which the same task was undertaken. Pupils also complete surveys that gauge attitudes and motivation.

A wide audience is reached by using a variety of reporting methods (online and in printed form). Those involved in education as well as the wider community have access to the project's findings. About two thirds of the tasks used in the project each year are made available for general classroom use. Equipment in packs is available for schools once the tasks are released. The remaining third of tasks are reserved for the next cycle of monitoring, so that performance can be compared from one cycle to the next.

#### **2.1.4 Considerations**

1. The survey includes tasks which enable practical skills to be assessed. This is particularly valuable in a practical subject such as science.

2. The tasks are short, manageable and engaging for pupils. Teacher-administrators find them straightforward to mark, simply recording which of a number of possible responses a pupil/group gives.
3. Teachers who train as administrators value highly the professional development they receive. Many administrators report advantages gained from seeing a variety of assessment methods, working with another teacher in other schools with small numbers of pupils.
4. Researchers watching videos of the assessments have noticed that the teacher-administrators vary in the amount of assistance they give pupils.
5. The training of administrators and production of materials involves significant expenditure.
6. Pupils seem to be benefitting from the project less than the teacher administrators: an increasing number of pupils say in the surveys that they ‘never’ do practical investigations in science. This may be because science is only assessed once every four years.
7. Some problems have arisen with the random sampling of pupils: selected pupils who are recent immigrants or disabled are often replaced.

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

### **2.2.1 Overview**

The National Center for Education Statistics (part of the U.S. Department of Education and the Institute of Education Sciences) is responsible for the production of the NAEP assessments. Educational Testing Service (ETS) is the primary NAEP contractor, with Pearson and some other organisations playing a smaller role. Tests are carried out with pupils in Grade 4 (aged 9-10) and also with pupils in Grades 8 and 12. It is the assessments for Grade 4 pupils that are considered in this case study.

Information on pupil performance is reported at national and state level and also for some large urban areas. Results are broken down by pupil groups, including gender, SES, ethnicity. Trends over time are reported.

### **2.2.2 Purpose of the assessment**

The National Assessment of Educational Progress (NAEP) was established in 1969 to monitor attainment over time in reading and mathematics (the Long-term Trend, or LTT, NAEP). A separate suite of tests was later introduced to enable state-by-state comparisons of attainment. This second set of tests became known as the main NAEP tests. Science is assessed as part of main NAEP. The tests, therefore, serve two main purposes: to track changes in national standards over time, and to compare achievement across states.

### 2.2.3 Assessment aspects

Less than four per cent of the total Grade 4 cohort takes the NAEP assessment. A sample of schools is selected and participation of these schools must be greater than 85 per cent for the results to be published. Within a selected school, 25 to 30 pupils are randomly chosen for each subject tested. Science is assessed on a four-year cycle and participation is voluntary.

The Grade 4 assessments are based on the science framework which breaks the subject down into the fields of science (physical sciences (33%), earth and space (33%), life sciences (33%)); and scientific practices (identifying science principles (30%), using science principles (30%), conducting scientific enquiry (30%), employing technological design (10%)) (Crovo and Raizen, 2005). The target amount of assessment time for each of these areas is in brackets. All items are pre-tested prior to live usage.

All sampled pupils take a version of the pencil and paper tests. The items are a mixture of multiple choice, short constructed-response and extended constructed-response, and span a range of difficulty. There are some overlapping items between the Grade 4 and Grade 8 assessments. The tests also collect background information from pupils, teachers and schools. The assessment takes 50 minutes and a further 10 minutes is allowed to complete a questionnaire. Approximately half of the pupils sampled to take the science assessment also take a hands-on task that lasts less than 30 minutes. NAEP staff administer the tests in schools and are responsible for overseeing the hands-on tasks. In 2009, a small number of Interactive Computer Tasks (ICT) were also used for the first time (Bennett *et al*, 2008).

For hands-on tasks, pupils are provided with a task and some objects with which to solve it. Pupils are assessed on both the solution and the procedures used to carry out the investigation. In 2009 there were four types of ICTs: information search and analysis, empirical investigation, simulation, and concept maps. Information search and analysis items posed a scientific problem. Data was provided for pupils to analyse and decide on relevance in order to address the problem. Empirical investigation items took the form of hands-on performance tasks on the computer. Simulation items modelled systems (e.g. food webs) and asked pupils to manipulate variables, and predict and explain resulting changes in the system. Concept map items provided pupils with terms which they had to link using arrows which they also label (National Assessment Governing Board, 2008).

Multiple choice items are electronically scanned and scored. Extended response items are scanned and then scored online. Hands-on tasks are provided to pupils in the form of a work sheet. The new computerised assessments are carried out on laptops taken into schools and actions and responses are recorded directly to the computer.

Results are given in the form of ‘the Nation’s Report Card’. Achievement levels are reported as basic, proficient and advanced and the particular skills associated with each achievement level are also reported. Success on individual items and what they aimed to assess are also reported. Audiences for the results include educators, parents, policy makers and the media.

## **2.2.4 Considerations**

1. The Interactive Computer Tasks (ICT) are hands-on and engaging for pupils.
2. Hands-on tasks enable a broader testing of scientific inquiry and skills which are closer to classroom experiences and reflect the importance of ‘doing’ science.
3. Challenges of this method of assessment are the time to carry out the assessments, the difficulty of creating tasks that are not just following directions, equipment costs, and scoring costs.
4. ICTs can be used to do things not possible in other formats, such as show processes in slow motion and assess skills in finding information; they are potentially cost effective; and can be used to simulate experiments that require a large number of readings, such as rolling a ball down different height slopes with different surfaces.
5. Concerns of using computer-based assessments include the availability of computers, limited research into computer-based assessment (although this is changing with the increased use of computer-based assessments, such as PISA electronic reading assessment in 2009) and initial development costs.

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

### **2.3.1 Overview**

The Scottish Survey of Achievement (SSA) was introduced in 2005 to replace the Assessment of Achievement Programme (AAP) which had been running since 1983. The SSA is run by the Scottish Government, in partnership with the Scottish Qualifications Authority (SQA) and Learning and Teaching Scotland (LTS).

### **2.3.2 Purpose of the assessment**

Both the SSA and the AAP before it shared the primary purpose of producing national estimates of achievement, for the entire pupil population at a stage, and for subgroups within this (e.g. boys/girls), and for monitoring standards over time. The SSA had the additional objective of doing the same thing for local authorities, where these were nominated for separate reporting (as in 2005 and 2006) or volunteered for it (2007 science and 2008 maths/numeracy). Additional main purposes were to provide a contextual background against which to view the achievement findings, in the form of information about learning circumstances and pupils’ subject attitudes gathered through questionnaires. CPD has been an important but subsidiary objective. It has not been compulsory for schools or pupils to take part in surveys, and no information is produced about individual pupils, teachers or schools.

The subjects assessed in the first five years of the SSA were English, maths, science and social subjects, one or other subject per year. The SSA is currently being revised and in the future will assess literacy and numeracy in alternate years.

### 2.3.3 Assessment aspects

The programme has focused on four year groups: primary 3, 5, 7 (ages 8, 10, 12) and secondary 2 (age 14). Science was assessed in 2007, having previously been assessed within the AAP in 1987, 1990, 1993, 1996, 1999 and 2003. In 2007, the achieved sample was 30,000 across all four age groups (7,000 to 8,000 per year group), this large sample size being explained by the need to produce separate attainment estimates for most of the Scottish local authorities. In 2003, the sample was less than half this size, as the reporting was at national level only. All pupils in the main sample took the paper and pencil tests, very much smaller subsamples of pupils in subsamples of the survey schools took practical assessments of one kind or another.

Surveys, which took place during a five week period in May/June, assessed attainment with reference to the 5 – 14 progressive level framework included in the national curriculum guidelines for the relevant subject. Written tests of science knowledge and understanding formed the core SSA in 2007, other forms of assessment playing a lesser role, including a small-scale exercise featuring investigation skills in science. Pupil and teacher questionnaires were used to gather contextual information about pupils' learning environment and experiences.

A team of 160 field officers (teachers nominated by the local authorities and trained) conducted or monitored the practical assessments. They worked in pairs visiting 4-10 schools each. Field officers used rating schemes or checklists to record pupils' actions and responses, and, where required, to come to decisions about attainment levels.

There were four kinds of practical assessment, with 3-4 pupils in any 'practical' school randomly assigned to one or other type:

- 1 Scientific Investigation Skills: Class teachers chose one practical investigation from a list of four (or used one of their own, as long as it complied with given criteria). The class teacher could use the investigation with their whole class or with just the randomly assigned pupils. The pupils discussed what they did one to one with a visiting field officer at a later date, using notes they made at the time of the practical investigation as a basis for the discussion.
- 2 ICT skills, assessed in a science context: Field officers observed individual pupils doing a 'virtual investigation' using ICT.
- 3 Working with others, assessed in a science context: Group discussions where pupils were given a scientific/ethical issue (e.g. what is important in designing a pair of trainers?) to discuss. The field officer observed and assessed one of the pupils in the discussion group, and also assessed the performance of the whole group.
- 4 Science Literacy: Field officers discussed topical scientific/ethical issues (e.g. Are zoos cruel to animals?) with individual pupils.

National reports are given to all schools each year (and are available online).

### **2.3.4 Considerations**

- 1 In the recent past the core SSA has used paper and pencil tests to assess across the curriculum in particular subject areas, including science. In 2007 the tests were designed to assess pupils' science knowledge and understanding. The paper and pencil testing was supplemented by very much smaller scale practical assessments, whose results were intended to be indicative of pupils' achievements, providing additional information on specific areas, and developing assessment expertise in the teaching profession.
- 2 Teachers who served as field officers felt strongly that they benefited from CPD.
- 3 Rating schemes were used by field officers for levelling pupils for some types of practical assessment; with training provided to ensure as far as possible that the field officers interpreted the schemes consistently.
- 4 For the informal assessment of science investigation skills that featured in the 2007 survey, field officers did not actually see the pupils doing the practical investigations; but discussed these with the pupils when they visited the schools at a later date.
- 5 Teacher assessments are also collected for the group of pupils being tested. Disparities are seen between the test results and the teacher judgements, especially in science (see Johnson and Munro, 2008).
- 6 Teachers are involved in the process, including for marking some aspects of the work, for assessing certain skills, and for providing teacher assessments based on class work. This capacity building within the teaching community can be seen as a very useful additional benefit of introducing national monitoring surveys.

## **2.4 Case Study 4: The Assessment of Performance Unit (APU) in England**

### **2.4.1 Overview**

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 Department of Education and Science (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.

### **2.4.2 Purpose of the assessment**

Initial purposes of the survey programme were to monitor standards over time for 11, 13 and 15 year olds, and to identify underachievement in particular pupil subgroups. 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 pupil subgroups. Monitoring of standards over time took a lower profile.

### 2.4.3 Assessment aspects

For the science surveys, samples of around 12,000 pupils per age group (about 1.5% of the population) were randomly selected for the pencil and paper testing, and all pupils also completed a questionnaire about their science learning experiences and views (see APU, 1988, Johnson, 1989, Foxman *et al.*, 1991, for details). Smaller subsamples of pupils, fewer than 1,000 per stage, were engaged in exploratory one-to-one practical investigations. The surveys were specifically designed not to be able to identify individual pupils, teachers or schools.

Pupils aged 11, 13 and 15 were tested at different times, the 15 year olds were tested in November and the primary pupils were tested in May/ June.

The pencil and paper tests were created and administered using a multiple matrix sampling strategy. That is, relatively large numbers of test questions were randomly selected from within question pools and distributed randomly across a number of test booklets. The test booklets were then allocated at random to pupils, so that ultimately every test booklet would be attempted by similarly representative samples of pupils. The large number of questions ensured that the science curriculum could be assessed as comprehensively as possible. The time taken for the 11 year olds to complete the paper tests was 45 minutes. The paper tests were administered by class teachers and marked by practising teachers, each booklet split between two individuals (APU 1988, Appendix 7, and Johnson 1989, Chapter 7).

In one model, which did not form part of the monitoring survey proper, the practical tasks involved individual pupils being given a problem to investigate and developing and carrying out a practical investigation. The tasks were administered by a small group of trained teachers (field officers), and took from 50 minutes to 1 hour for 11 year olds to complete. During the course of the investigations the individual pupils were observed by a field officer and were asked questions at the end. The observations were recorded on check lists. These individually administered practical investigations were few in number, were used with relatively small numbers of pupils, and showed a high degree of variation in pupil performance from one task to another. The second type of practical assessment was conducted on a larger scale – the ‘circus’ assessment of observation skills, in which pupils moved from one ‘station’ to another attempting various tasks.

In the APU, reporting was at the country level (England, Wales and Northern Ireland) and the national level. As mentioned above there was no reporting of results for individual schools, teachers or pupils.

Reporting was based on different sets of skills or areas of the curriculum, rather than a single overall score. Patterns of errors in different clusters of items were reported. The emphasis was on comparing the performance of subgroups of pupils at particular points in time, such as 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. Although, monitoring of standards did feature (e.g. APU, 1988, chapter 4).

### **Considerations**

- 1 The small scale, one to one assessment of practical investigation skills in science, while ground-breaking, was time consuming and expensive. In addition, training for assessors was required. The ‘circus’ model was less time consuming.
- 2 Pupils were assessed individually for the ‘performing investigations’ practical component.
- 3 The administration of the tests with only seven pupils in each school made them logistically difficult to manage.

## ***Teacher Assessment Case Studies***

### **2.5 Case Study 5: Scientific Minds – A guide to assessing attainment target one**

#### **2.5.1 Overview**

*Scientific Minds* is a publicly available teacher resource produced by nferNelson in 2005, developed by NFER. The *Scientific Minds* series consists of three books – Key Stage 1, Key Stage 2, and Key Stage 3. For the purposes of this case study only the Key Stage 2 edition was considered (Jones and Griffin, 2005).

The book describes techniques that teachers can use to assess pupils’ performance in Attainment Target 1 (AT1). The book contains four different activities. Each activity consists of a list of materials needed, a description of the investigation, links to the National Curriculum Programme of Study, provides guidance on how to assess pupils’ performance and how to level pupil work. The aim of the book is to give teachers ideas about how assessment of AT1 can be carried out, and to help teachers use the results of assessment meaningfully.

#### **2.5.2 Purpose of the Assessment**

*Scientific Minds* provides exemplars of activities teachers can use in order to carry out teacher assessment.

#### **2.5.3 Assessment aspects**

Pupils carry out an activity selected by their teacher. Worksheets are provided in the book and during the course of the activity pupils complete the worksheets. The guidance provided in the book identifies the areas which could potentially be assessed by carrying out the selected activity and the worksheets are tailored to this. Teachers are then responsible for marking the pupils’ work and assigning levels.

The activities are designed to be used with whole classes, where a written record can be marked. The activities may also be used with one pupil or a small group where a teacher observes the pupils while they work and asks them questions.

There is no indication of the frequency of the assessment but there are only four activities. The book states, 'After you have used the activities in this book, you should find that you are able to see opportunities for assessing AT1 during other class activities, with little or no adaptation of tasks or interruption of learning.' (Jones and Griffin, 2005, p. 1). The length of the activities varies depending on how the teacher decides to use them. Estimates of the length of the activities are provided in the book, and range from one lesson to three one-hour lessons.

Within the book detailed descriptions of the National Curriculum levels are provided, linked to the particular Programme of Study reference each activity aims to assess. The book also provides examples of pupils' work in 'pupil speak' at various levels. This provides primary teachers (science specialists and non-specialists) with the tools to assist in levelling pupils' work.

The levels assigned to particular activities can be used to contribute to an overall teacher assessment (TA) level within AT1. More evidence from a wider range of work is needed to build up a definite judgement of a pupil's level in AT1. The level achieved in AT1 may then be used with the National Curriculum levels from other areas of the science curriculum to provide an overall level in science, which may then be reported.

#### **2.5.4 Considerations**

- 1 The assessment lists the equipment required for each task.
- 2 A summary is provided for teachers with regards to length of activity, the purpose of the activity, which parts of the Programme of Study are being assessed etc. This provides a quick way for teachers to determine whether an activity is appropriate for a particular area of content/skills.
- 3 Examples of 'pupil speak' answers are provided which make it clear what a teacher should be looking for.
- 4 Pre-prepared worksheets which are specifically designed to assess particular areas of the curriculum, such as planning, are also included.
- 5 The tasks are challenging and the ways in which pupils present their work innovative and interesting.
- 6 Practical activities are trialled in order to ensure they work and any equipment needed is accessible to primary teachers.

## **2.6 Case Study 6: Year 4 science optional tasks**

### **2.6.1 Overview**

This series of assessments were designed to assess pupils halfway through Key Stage 2 (at the end of Year 4). They were written to assess English, mathematics and science. The focus of this

case study will be the science materials. Five units were developed and published in late 1997 to assess pupils working at Levels 2-4 across the Key Stage 1 and Key Stage 2 Programmes of Study. The units are named *Humans*, *Plants*, *Materials in the Environment*, *Forces and Motion*, and *Light and Electricity*. Each unit has two separate activities. Only one set of these materials was produced for schools. These materials were developed by NFER and published by QCA in 1997.

### **2.6.2 Purpose of the assessment**

The materials were '*designed to assist (teachers) in making accurate judgments about children's achievement halfway through Key Stage 2 and to support schools in planning effectively for the second half of Key Stage 2*' (QCA, 1997, p3). The evidence gained in these assessments is designed to complement the teachers' own knowledge of their pupils' performance rather than form the entire evidence base for assessment.

### **2.6.3 Assessment aspects**

The Teacher's Guide states that the materials can be used 'during Year 4'. Teachers can use any part or all of the assessments in each unit. They can be used as part of a unit of teaching or at the beginning and/or end of a teaching unit. The assessments can be done with the whole class, small groups or individually.

The assessments are designed to allow the teachers to interact with the pupils during the assessments to allow further clarification of pupils' responses if needed. This is easier to achieve in practical terms with small groups or individuals rather than bigger groups. The individual task booklets suggest times that most pupils will take to complete each task.

The Teacher's Guide states that evidence may be gathered on other elements of the National Curriculum, such as in '*Speaking and Listening, Using and Applying mathematics and Experimental and Investigative science*' (p4). How the assessments contribute to this information is not provided.

In preparation for each task, teachers are given an overview of the context, which elements of the Programme of Study are assessed and what is needed for the task. Each task has a table indicating '*what to do*', '*what to say*', '*what to look for*' and '*example responses*'. There are photocopiable worksheets for pupils for some of the activities. To translate the responses into levels, there is a table indicating the responses and understanding that is typical of each level in that task. Exemplars of pupils' work at different levels are provided to support teacher judgements.

The types of tasks included in the assessments are often very practically based, such as drawing the position of shadows of sticks and predicting how the shadows will change throughout the day, building a circuit including a light bulb, measuring foot length and determining if there is a relationship with leg length, and observing what happens to a puddle over time. These tasks

encourage the pupils to do real experiments and investigations indicating that this type of assessment has construct validity. As these are school based activities, any equipment needed is provided by the school.

During the development, teacher questionnaires filled in at the time of trialling suggested that 71 - 85% of teachers who trialled the materials would be 'likely' or 'very likely' to use the materials if they became available.

#### **2.6.4 Considerations**

- 1 There is information contained within the Teacher's Guide to support the administration of the materials and to help teachers interpret the results with respect to level. Exemplar materials for pupils performing at different levels are included.
- 2 The flexibility of the assessment is a strength, so that teachers can judge when pupils are likely to perform at their best.
- 3 Teachers are able to pick and choose which activities they want to use.
- 4 Each activity clearly sets out what the teacher needs to do, what prompts to give where necessary, what answers are expected and some example responses. Worksheets are given for some of the activities.
- 5 The materials only partially assess the curriculum, so teachers should/could not rely on these as the sole means of assessment.
- 6 Teachers have the option to ask for clarification of responses given by pupils. This may be more difficult to manage if the assessment is taking place in larger groups rather than individually.
- 7 The variety of activities, including practical investigations, is a strength of this set of materials.

## **2.7 Case Study 7: Assessing Progress in Science**

### **2.7.1 Overview**

This set of materials is designed to encourage teachers to incorporate assessment into the teaching and learning cycle. If teachers are able to gather information on what has been learnt, this information can be used that to adapt future lessons. The assessment aims to provide diagnostic information to teachers and pupils.

The materials cover a broad sweep of the curriculum at Key Stage 2, but they do not assess the entire curriculum. There are eight units in Key Stage 2 assessing elements of Sc1 - Sc4. The assessment activities themselves each have a teaching and learning sequence followed by a number of assessment activities. Notes are provided on reviewing and interpreting the evidence presented by the pupils.

The materials were produced in 2003 by QCA and were written by researchers at the Centre for Research into Primary Science and Technology (CRIPSAT).

## 2.7.2 Purpose of the assessment

*Assessing Progress in Science* aims to provide activities for teachers to use to formatively assess their pupils across a range of topics. It gives teachers a model to enable them to develop formative assessment activities in other areas of the curriculum.

## 2.7.3 Assessment aspects

The units are not designed to be used in any particular order; the guidance is to select tasks according to schools' own schemes of work and the abilities of the pupils involved with due regard to health and safety implications (Russell and McGuigan, 2003).

As with other teacher assessed activities discussed in this paper, teachers are able to encourage pupils to express themselves fully during the assessment. Photocopiable 'ideas sheets' have been produced for most of the units with the aim of encouraging pupils to 'describe their ideas fully'. Teachers can refer to these sheets when asking for fuller explanations. The units may be worked through on a one to one basis, in pairs, in small groups or as a whole class.

The activities are quite diverse across the range of topics, including classifying plants and animals, making observations on different materials, separating mixtures and building a circuit as part of an investigation.

As with the previous case study on the Year 4 Tasks, there are some examples of expected pupil responses to help interpret the pupils' work.

## 2.7.4 Considerations

- 1 The teacher's guide is quite theoretical; this may be helpful if the aim is to increase teachers' knowledge and understanding of the different types of assessment.
- 2 Providing ideas sheets may help pupils to pin down any ideas they have and the teacher has a record of the original thoughts, if they want to ask the pupil to clarify any of their ideas. However, having a very open response ideas sheet may prove daunting to some pupils where more scaffolding may help to elicit ideas better. Across the range of topics, some sheets have more scaffolding than others.
- 3 Examples of answers that pupils tend to give are provided as are types of response given by pupils working at different levels.
- 4 Along with methods of interpreting answers with respect to the National Curriculum, there is some advice on feeding back to pupils. This guidance tends to be very general and it is difficult to know whether teachers would find this type of advice useful. There is also guidance on how to use the information in teaching and learning but this tends to give more activities that could be used to teach the same part of the Programme of Study, rather than how to tailor teaching and learning to any outcome of the assessment activity.

- 5 The activities are not explicitly linked to the Programme of Study. Where references are given, it is to the Schemes of Work. When describing performance at a level, there are some references to the level descriptors.
- 6 Examples of pupil responses in the level charts are not given in ‘pupil speak’. Giving examples in ‘pupil speak’ may help teachers interpret responses more reliably.
- 7 The activities are manageable for teacher assessment in terms of time and resources. Teachers will need to spend some time interpreting responses, feeding back to pupils and incorporating data into the teaching and learning cycle.
- 8 There is an appropriate range of activities across the science curriculum, including practical and written components.

## **2.8 Teacher Voice Omnibus Survey (June 2009)**

NFER runs a teacher omnibus survey with a panel of 1000 teachers. In June 2009 a number of questions were included that were related to the assessment of science. Primary teachers were asked about the impact of the loss of the Key Stage 2 science National Curriculum tests and their preferences as to how to assess science at Key Stage 2 in the future. The majority of teachers felt that the removal of the tests would produce an improvement in pupils’ learning and more reliable reporting of science results at Key Stage 2. Thirty five per cent of respondents thought that it would cause an increase in teachers’ workload. In terms of replacing the tests, the two most popular options were teacher-set or externally-set tasks involving planning and carrying out investigations with teacher marking, with about 80% of respondents choosing one of these preferences. The next two most common choices were teacher-set or optional externally-set tests with teacher marking which were chosen by 50% of respondents. These responses suggest that teachers value marking assessments themselves possibly because of the extra information they gain from the process. There were a number of open response comments on assessment suggesting that teachers feel there may be some need for assessments like the Key Stage 2 tests to be provided for new or weaker teachers as they help establish some idea of the standards expected at that age range, although many felt that written tests are not a good assessment of science.

## 3 Discussion

As with any assessment system, the purpose(s) need to be clearly outlined in the initial stages, as many of the decisions with respect to structure, administration, marking and reporting relate back to the purpose of the assessment. In this paper, the assessment of Key Stage 2 science is outlined with two completely separate purposes: national monitoring and teacher assessment. These will be discussed separately in Sections 3.1 and 3.2 below.

The case studies referring to national monitoring are the National Education Monitoring Project (NEMP) in New Zealand, the National Assessment of Educational Progress (NAEP) in the USA, the Scottish Survey of Achievement (SSA) and the Assessment of Performance Unit (APU) in England.

The case studies referring to teacher assessment are *Scientific Minds: a Guide to Assessing Attainment Target One* (UK) (Jones and Griffin, 2005), *Year 4 Science Optional Tasks* (UK) and *Assessing Progress in Science* (UK) (Russell and McGuigan, 2003).

The NFER Teacher Voice Omnibus survey (June 2009) asked teachers about their preferences for assessing science in the future and the impact of no longer having the end of key stage tests.

### 3.1 National monitoring

In Ed Balls' response to the Expert Group on Assessment in May 2009 (DCSF, 2009), there was a proposal to introduce a national monitoring system at the end of Key Stage 2 in science. The stated purpose of the national monitoring or sample tests would be to externally validate national standards in science but not to report back to schools or local authorities. It should be noted that a national monitoring survey is likely to be low stakes to the schools and pupils and as such it will be difficult to map performance in this back to Key Stage 2 levels. It would be more straightforward to monitor standards over time, from when the survey was introduced.

The National Curriculum tests developed for 2010 and 2011 will most likely form the initial monitoring tests, but the aim is to design a purpose built assessment instrument for 2012. This paper discusses some possibilities for the design of this new monitoring assessment in terms of sampling, structure, administration and manageability, marking and reporting, and requirements for CPD based on lessons learnt from the case studies described above.

#### 3.1.1 Purpose(s)

As stated previously, it is essential that the purpose is clearly defined at the outset. While external validation of standards is one purpose of the monitoring assessment, another is to track standards over time (NFER, 2009, p. 5; DSCF, 2009), or to monitor strengths and weaknesses in different curriculum areas or among different groups of pupils as in the APU.

Recommendations from NFER's recent report *Submission to Expert Group: Issues to Consider When Developing a National Monitoring System* (NFER, 2009), state that the main purposes of a monitoring system for Key Stage 3 science should be two-fold: monitoring changes to absolute standards over time and investigating areas of strength and weakness across the curriculum. One other aspect that could be monitored is the performance of various sub-groups, although sampling would need to be designed to take this into account. These purposes will also apply to monitoring at Key Stage 2. As long as the sample design and assessments take it into account, it may also be possible to monitor development in some aspects of science across Key Stages 2 and 3. To do this, there would need to be common items across the two key stages and a careful analysis would need to be conducted on the curriculum at the two stages. Common items could be chosen from those aspects of science accessible at both key stages.

In terms of purpose, is it necessary to have a practical component to the monitoring? One of the perceived limitations of the National Curriculum science tests centres around the issue that science has a practical component and this is not assessed in a paper and pencil test. All the case studies reviewed in this paper suggest that a practical component can be used effectively to fully assess science, although this will impact on the cost of the system.

It will be important to make explicit what the monitoring survey measures. If it is a pencil and paper survey, it may not measure the same areas of the Programme of Study as teacher assessment and therefore the results may be different and must be taken in conjunction with teacher assessment (depending on the way in which the results will be used). If it is to act to support teacher assessment, the structure of the assessment needs to ensure that it measures something as close to teacher assessment as possible, with a practical, written and possibly other components.

National monitoring in science will allow the government to monitor any change in performance over time that may result from no longer using the end of Key Stage 2 National Curriculum tests in science. Many in the science education community, such as the ASE, the Royal Society and SCORE welcomed the cancellation of the Key Stage 2 science tests and the chance for schools to broaden and adapt their lessons. There is also some concern that as science is no longer a 'core' subject in the new Key Stage 2 curriculum, it may suffer from a drop in emphasis in schools. It will be useful to monitor that this does not happen.

In NEMP, the main purpose is to get a broad picture of achievement nationally over time, to report trends for public information and to inform policy making, curriculum development and educational planning. A recognised benefit or secondary purpose is to help with professional development of teachers who get involved with the administration. In APU, the main purpose of the assessment was to determine the strengths and weaknesses within subjects and in different subsets of the population. The monitoring of standards over time was secondary to this purpose. In NAEP, the main purpose is to monitor pupil achievement across the different elements of science. Performance at the item level is reported, as well as proportions of pupils reaching different levels of achievement. These examples demonstrate clearly that the detail and nature of

the agreed purposes and the type of reporting will have a significant impact on the design of the tests. In SSA, one key benefit is the development of assessment expertise in the teacher workforce, achieved by the inclusion of teachers in the administration and marking of the assessments.

### **3.1.2 Structure**

This relates to a number of elements of the assessment system including: the type of test (pencil and paper/e-assessment/practical); how many different elements of the test there are and the number of marks; what it measures and how the different components are weighted; and whether the entire curriculum is assessed each time the assessment occurs (either one long assessment or in a number of test versions) or whether the curriculum is assessed over a number of assessment cycles where each cycle has a specified focus. Many of these elements are interdependent and are related to purpose. Any monitoring system needs to ensure that judgments made by the assessors are consistent and reliable across the cohort being measured and from year to year. This will need to be considered within the structure. Cost and manageability will also impact on the final structure.

Running a practical assessment alongside a pencil and paper or computer-based assessment will increase the costs considerably, although it is likely that a practical assessment would be perceived more favourably within the science community as being a better assessment of science. Whether to have a pencil and paper based assessment or a computer-based assessment is another choice. A pencil and paper based assessment is likely to be more cost effective to produce but more expensive to mark and moderate. A computer-based assessment could be more expensive to produce but marking and moderating could be relatively straightforward. Within this latter option, a simple multiple choice computer-based assessment would be less expensive to set up than a model using more complex, virtual investigations and/or open response items, but is likely to have far less credibility within the science education community. This is an area that needs further discussion, but is ultimately likely to be a policy decision based on considerations of validity and what can be achieved within the given budget.

Sample size can limit the ability to look at trends across the curriculum over time. Assessing part of the curriculum each cycle will have cost advantages, but it will mean that it is not possible to ‘take a snapshot’ of the performance of any one cohort against the whole Programme of Study. Monitoring part of the curriculum, rather than most [all] of it, would require more frequent testing. For assessing most of the curriculum, the options are one long test or a number of shorter papers. The sampling implications of each of these choices are discussed in the section on sampling below.

### **3.1.3 Sampling**

In any sampling design, there needs to be a balance between ensuring that the data is sufficiently precise to produce valid findings and schools not being over burdened. This is likely to be less of an issue for monitoring at Key Stage 2 rather than Key Stage 3, due to the greater number of

primary schools than secondary schools in England. The Expert Group on Assessment recommended that it should be compulsory for schools to participate in any national monitoring surveys, reflecting a recommendation that had been made in the NFER submission to the group. This may help to alleviate some of the difficulties agencies have had in recent years in achieving samples for surveys such as TIMSS, PISA and PIRLS, and even in pre-testing National Curriculum papers. In making these surveys compulsory for schools drawn in a sample, however, it does not mean that it should be difficult for schools to take part. Having schools accept that it is part of their role is likely to mean that pupils take the survey more seriously. Surveys such as NAEP and SSA have also had difficulty in achieving the specified sample.

The stated purpose of national monitoring is *'to track national standards, enabling the Government, educational professionals and the public to see the progress over time of pupils, and the effectiveness of education policies'* (Bevan *et al.*, 2009, p. 32). The same report indicates that any data from national monitoring would not be used for school or LA accountability. The tests would, as a result, be low stakes for both schools and pupils and this may affect performance. There is evidence that performance on low stakes tests is significantly different to performance on high stakes tests (Wise and DeMars, 2005), and that the effect may be more complicated than just an underestimation of overall performance (Pyle *et al.*, 2009). This would need to be taken into account if comparisons are made back to National Curriculum performance prior to 2010.

In NEMP (New Zealand), approximately 3 per cent of the population in any one year group are sampled in around 260 schools. The schools are classified according to area and school type and then randomly selected. Participation is not compulsory and schools may withdraw for a number of reasons. In NAEP (USA), about 100 schools are selected in each state for each subject tested in each grade, which is about one per cent of the public school pupils in each grade being assessed. The sample is stratified on the basis of *'physical location of the school, extent of minority enrolment, state-based achievement scores, and median income of the area in which the school is located'* (NCES, 2009, front page). About half the pupils in each school do a practical 'hands on' activity. In the SSA (Scotland), the sample consists of around 4,000 per year group. In 2007, this was boosted to 7,000 to 8,000 to allow reporting at the local authority level. Pupils were randomly selected from about 1,300 selected schools, with a small number of pupils required to take part in each school. Practical assessments were taken by a subset of this sample: up to three pupils in each year group in each practical school for each of the practical assessments. The sample design for APU was similar to SSA: there were a large number of schools taking part with a small number of pupils in each. Seven pupils from each school took part and a total of 10,000 pupils did the written survey and a subset (2,000-3,000) did the practical element of the survey.

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.

### **3.1.4 Administration and Manageability**

One of the main criteria for any national monitoring system is that any judgements made are as consistent as possible. For a paper and pencil test, checking is needed during marking to ensure the appropriate consistency. In a computer-based multiple choice test, the computer will apply the mark scheme and the results would be very consistent. In a computer-based assessment using a more open response model, the mark schemes will be more complex, but there would be an element of consistency in how the mark scheme is applied. Both pencil and paper and computer-based models are easy to administer to a relatively large sample and it may be possible for schools to administer these tests themselves, especially if there is no accountability data produced on school level performance.

The time required for each pupil to participate in the national monitoring exercise could impact on how it is viewed in each school; obviously the longer the time that is required, the more likely schools are to view it in a negative light although this may be mitigated somewhat if a high level of support is provided alongside the national sampling administration.

Practical assessments are more complex to administer, requiring observation and the recording of results fairly and consistently. The various examples looked at in the case studies have managed this in very different ways.

NEMP has a practical element that is taken into schools by administrators. Administrators are teachers that have been seconded from their own classrooms for a period of approximately five weeks to allow for training and the administration. In NEMP, there are four types of practical administration, although of the tasks released, only three have been used in science. They are group work (four pupils working together co-operatively), one to one (pupil working with the administrator) and a stations approach (four pupils working independently around a series of stations). The tasks involve each pupil working with the administrator for between 3 - 4 hours over a period of five days. The tasks themselves are short and easy to administer, the instructions are clear and concise. The tasks are often recorded on video to allow them to be referenced at a later date if needed. This approach requires administrators to be in each school for about a week.

In NAEP, about half the pupils in each school do an individual practical activity. The example activities that have been released are very structured, and the tasks tend to assess the science content rather than looking at the investigative parts of the activity. Answers are recorded in an answer booklet.

In 2007, the SSA used an assessment model for the supplementary practical activities in which an investigative task was used either with the three target pupils, or with a whole class group. Field officers then visited the school to look at and discuss the results with teachers and pupils. This is an unusual way of assessing the practical element of science, although it may be a more

manageable and cost effective way of assessing practical skills and provides additional information to simply using paper and pencil methods.

For APU, practical tasks were assessed using a subset of the main sample. They were administered by a small group of trained teachers and observations were recorded throughout the process. The practical tasks were carried out on a one to one basis.

In each of the examples above, administrators or field officers were used for the practical elements, which helps ensure consistency of administration and interpretation of data. A secondary benefit of this approach is the professional development of the teachers who are trained to be administrators in any assessment cycle. They gain experience of pupils outside their school, in moderation procedures and more assessment expertise generally. Teachers from NEMP, for example, cite this as a real benefit of being an administrator. The obvious disadvantages of relying on administrators are that they are only needed for a relatively short period of time and taking them out of the classroom could cause staffing issues in their schools. Again from NEMP, the administrators work in pairs and each pair assesses five schools in five weeks. In each cycle, 260 schools are drawn in the sample. For one cycle, 104 teachers are required for five weeks and schools would need support in dealing with this issue.

The type of activity used in the practical tasks and what is being assessed will also determine the number of pupils that can be assessed at any one time. The more complex the task, the more difficult it would be to determine who in the group was driving the task forward and contributing most to the task. If assessing pupils individually, skills such as those required in a group work situation cannot be assessed but it is easier to determine what contribution each pupil makes. NEMP gets around this by assessing pupils in a variety of practical situations, but this approach is expensive, time consuming for the school and administrator-intensive.

### **3.1.5 Marking and Reporting**

Marking paper and pencil tests requires a group of trained markers. The model that has been applied in marking National Curriculum tests could be applied in this situation. The use of marking centres may help to speed up the process, but this would also have the disadvantage of excluding current teachers who are likely to benefit most from any professional development that occurs as a result.

For computer-based assessment, marking is likely to be much faster and largely automated. A smaller number of markers may be retained for extended-response or more complex answer types.

Practical observations require a different set of skills again. If the observation can be set up to allow observers to make simple observations as they go, as in a number of case studies described above, each observer can only assess a small number of pupils at one time. This ensures that pupils are given the opportunity to respond to questions orally or using other methods that are

not dependent on writing. NEMP for instance, allows pupils to present information orally, by modelling or computer output.

Using administrators helps to make the administrations and interpretation of the observations more consistent across the survey. Marking and interpreting observations have a number of issues, such as moderation and marker training, and making interpretations of those involved in group work activities. A real advantage of using a survey approach for monitoring standards over time is that it is not necessary for all pupils to sit the same test versions. A large number of questions can be included, with a subset of the pupils using a subset of the questions. This means that a larger proportion of the curriculum can be assessed in any administration. Which allows robust measures to be reported of performance in different areas of the curriculum, in addition to tracking broader standards.

## **3.2 Teacher Assessment**

Following the publication of *The Report of the Expert Group on Assessment* (Bevan *et al.*, 2009) it became clear that teacher assessment of science will need to be strengthened in primary schools. The Expert Group made it explicit that science should be assessed at Key Stage 2 through high quality teacher assessment supported by materials which help teachers to continue to improve their assessment skills.

From 2010, *Assessing Pupils' Progress (APP) in Science at Key Stage 2* (see <http://www.qcda.gov.uk/13581.aspx> for more information) will be available for primary teachers to aid teacher assessment. APP consists of a range of tools which aim to “support teachers in making robust, reliable and educationally useful assessments” of pupils’ work (Bevan *et al.*, 2009, p. 12). APP aims to provide a structured approach to assessment and contributes to the professional development of teachers.

The recommendations put forward here are for teacher assessment investigative tasks which would work alongside APP. APP can be thought of as one aspect of assessing science. High quality teacher assessment should encompass a range of activities and techniques. The recommendations given in this section are based on pupils carrying out robust practical activities which can then be assessed by teachers.

NFER has been conducting a research project investigating the introduction of APP in secondary schools. More information can be found at [www.nfer.ac.uk/nfer/research/projects/pupils-progress-science/pupils-progress-science\\_home.cfm](http://www.nfer.ac.uk/nfer/research/projects/pupils-progress-science/pupils-progress-science_home.cfm).

### **3.2.1 Purpose(s)**

High quality teacher assessments should be based on a variety of activities around practical investigative science which can be used flexibly to support teaching and learning.

### 3.2.2 Structure of Assessments for Teachers

In order for pupils to carry out practical tasks which can be assessed, teachers need to be given clear, concise and relevant guidance on how to carry out the tasks and how to assess the outcomes. Each activity should clearly state:

- approximately how long the practical task will take;
- links to the appropriate Programme of Study reference and APP Assessment Focus (what it is assessing);
- a list of the apparatus required;
- clear descriptions of how to set up and carry out the practical task;
- what prompts the teacher should give when necessary;
- what answers are expected from pupils;
- examples of pupils' responses;
- worksheets or clear details of what work to collect in order to assess pupils.

A summary of these details should be provided at the start of a practical task so teachers are able to determine whether the activity is appropriate for a particular purpose.

### 3.2.3 Administration and Manageability

Any practical tasks produced for teachers need to be thoroughly trialled in order to ensure they are manageable and produce suitable results across a range of classroom situations.

Ideally, practical tasks should be able to be administered to a whole class as a way of minimising manageability issues. Teachers should also be able to choose whether pupils work individually, in pairs or in groups on tasks.

The provision of pre-prepared worksheets specifically written for the area being assessed reduces teacher workload. Such worksheets also make marking and levelling more consistent and reduce the amount of time taken to mark. However, allowance should be made for teachers to adapt the materials to suit their own context if appropriate.

As with the Year 4 science optional tasks (case study 6) and Assessing Progress in Science (case study 7) opportunities for pupils to express themselves orally to clarify their understanding are considered to be beneficial. Guidance which describes opportunities for pupils to express themselves orally or use other methods (e.g. modelling) would be helpful for teachers.

Flexibility is a key component of successful teacher assessments. Materials which are flexible to administer and can be stopped and started to suit the teacher and/or pupils make assessment manageable. In addition, materials which can be used as a part of a unit of teaching or at the beginning and/or end of a teaching unit are helpful as teachers can choose to use whichever activities suit their plans (see case study 6, Year 4 science optional tasks).

### **3.2.4 Marking and Reporting**

When considering the marking of Teacher Assessments the main factor to consider is the ease with which consistent marking and levelling can be achieved. It is important to provide teachers with a clear and structured approach to assessment. The best way to achieve this is to provide information in a teachers' guide which helps teachers interpret results with respect to level. In *Scientific Minds* (Case Study 5), *Year 4 Optional Tasks* (case study 6) and *Assessing Progress in Science* (case Study 7) exemplar materials for pupils performing at different levels are included. It is also helpful if examples of pupil responses are given in 'pupil speak'. Giving examples in 'pupil speak' may help teachers interpret responses more reliably.

In *Assessing Progress in Science* (case study 7), as well as methods of interpreting answers with respect to the National Curriculum, there is some advice on feeding back to pupils. Within such advice it is possible to identify common misconceptions, and provide both pupils and teachers with further steps to help move the pupils forward in their learning.

If the results of teacher assessment are to be reported it may be appropriate to introduce some means of ensuring that standards are comparable across teachers and across schools. Moderation approaches could be introduced along the lines of those used at Key Stage 1. NFER recently held a seminar on methods for ensuring reliability of teacher assessment at Key Stage 3, and the outcomes are likely to be of use to this debate. The proceedings from the seminar can be found at <http://www.nfer.ac.uk/nfer/publications/policy-papers/>. A more detailed overview of approaches to moderation is given in the NFER paper on national monitoring (Maughan, 2009).

### **3.2.5 Other considerations**

By producing a teacher guide which is detailed and provides information about the purposes of the assessment tasks it is potentially possible to increase teachers' knowledge. Clear guidance on how to assess pupils' work consistently and reliably would contribute to teacher professional development and the professionalisation of teaching as a whole.

Any Teacher Assessments developed to be used in schools should not be designed to cover the whole curriculum. Instead, teachers should be provided with a few good examples of practical tasks for their pupils and should be given helpful guidance on how to level pupil work, with a view to developing their assessment skills.

## **3.3 Teacher Voice Omnibus survey (June 2009)**

Teachers on the panel for the NFER Teacher Voice Omnibus survey in June 2009 (see <http://www.nfer.ac.uk/what-we-offer/teacher-voice/> for more information) were asked about their preferences for assessing science in the future and the impact of no longer having the end-of-Key-Stage tests.

Primary teachers were supportive of the decision to remove the external tests for science at Key Stage 2 as 68% indicated that it will lead to an improvement in teaching and learning. The majority also think that Key Stage 2 results will be more reliable without the requirement for each pupil to sit the externally set test. A large proportion of the teachers surveyed indicated that they wanted to mark any assessments themselves, whether they were teacher set tasks or tests or optional, externally set tasks or tests. They commented that a range of assessments would provide better information about their pupils, but that any assessments did need to be moderated; one teacher indicated that the assessment should be based on *'teacher assessment of agreed knowledge, skills and attitudes with testing just being part of the assessment process. Teacher marking improves teacher knowledge and moderation'*. However, another teacher commented that external assessments do allow new teachers to develop an understanding of the levels expected.

## 4 Recommendations

### 4.1 National monitoring

The nature of a monitoring or sampling survey will depend on decisions that are made with respect to purpose and how the data will be used. These decisions will impact on all future decisions, and as such the following recommendations must be taken in the context of the initial discussions.

#### 4.1.1 Purposes

- 1 A significant part of the science curriculum is the practical/investigative element and the criticisms of the current tests focused largely on the absence of this element. A national sample monitoring system which includes a practical element would be a better assessment of the science domain. It also may encourage more practical investigative work in science teaching.
- 2 It is difficult to know how monitoring will impact on teaching and learning in classrooms; the monitoring tests will be lower stakes than the National Curriculum tests. If schools do not receive results monitoring may have little effect on school behaviour. However, as the National Curriculum tests were high stakes and adversely impacted teaching and learning, it is argued here that the monitoring tests should not attempt to influence what goes on in the classroom through the tests.

#### 4.1.2 Structure

- 3 When designing the structure of the assessment, methods to ensure the consistency of administration, marking and interpreting the data need to be built into the system. This may include, for example, using administrators in practical assessments.
- 4 Including a practical component will improve the validity of the survey as an assessment of science and will ensure that it meets the expectations of a wider group of stakeholders providing the sample numbers are fit for purpose. We therefore believe this will be an important component of the new assessment, although it will increase the costs of the running the monitoring survey.
- 5 Assessing most of the curriculum in each assessment cycle will provide a snapshot of performance in science. Assessing a focused part of the curriculum would mean that several cycles would have to contribute and a picture of performance at any one time could not be easily ascertained.

- 6 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.
- 7 If monitoring surveys across Key Stages 2 and 3 are designed in conjunction, it may be possible to look at development across the two key stages in some aspects of science.

### **4.1.3 Sampling**

- 8 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 assessments.
- 9 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 subset of pupils will be used for any science practical tests.

### **4.1.4 Administration and Manageability**

- 10 The burden on schools taking part in any national survey should not be onerous and, ideally, schools should have something in return for participating as recognition that they have done so. This does not necessarily have to be financial; rewards may be offered in terms of professional development for example. Any demands on school staff as a result of participation in the surveys should be kept to a minimum and should be designed to fit within school organisation as far as possible. Practical tasks do tend to be more burdensome both in the resources and organisation required to administer them, but they are seen to measure a component of science that was not measured by the National Curriculum assessments. As the development of practical skills is seen as valuable within the science community, this is an area that should be included in a national monitoring assessment. Teachers could be trained as administrators which would impact positively on classroom practice.
- 11 It is recommended that the use of technology be considered for both the administration of the tests and the marking. This could allow the use of more complex data in science, or simulations or slow motion video. It is also recommended that the background data on the pupils and attitudes to learning be collected in the form of an online survey.
- 12 Paper and pencil or computer-based assessments could possibly be administered by school staff. Trained administrators should be used for any practical assessments, for consistency across administration centres and in interpreting observations.
- 13 Some of the case studies (e.g. NEMP) involved assessment of group work. This has some advantages if group working skills are being assessed, although it is difficult to assess any one particular pupil in a group working environment. NEMP uses a mixture of group

working and individual working administrations to get around the issues associated with assessing individuals in group working. If group working were identified as an important part of the science curriculum, then a mixture of administration types may be more appropriate.

#### **4.1.5 Marking and Reporting**

- 14 Marking pencil and paper assessments and practical assessments requires trained markers and observers and/or moderation of the marking to ensure that there is consistency across the cohort. The design of practical administrations carried out by observers must consider how much any one observer can reliably do at one time; observations must either be simple and easily recorded or more complex and recorded in some way to allow the observer to replay it (e.g. video or audio recording). The number of pupils being observed will also impact on how the data is collected.
- 15 As with many aspects of this assessment, a clear definition of the purpose and what will be reported will determine the number of pupils that will need to be practically assessed. If the numbers of pupils are too small in any subset, it may not be possible to report valid information.
- 16 It is recommended that the survey design includes ways of linking the results to results from the TIMSS survey at grade 4 as this is the closest to Key Stage 2, to allow international comparisons to be made.

## **4.2 Teacher Assessment**

### **4.2.1 Purposes**

- 1 High quality teacher assessment tasks should provide a variety of activities based around practical investigative science which can be used flexibly to support teaching and learning. Teacher assessment tasks should not aim to cover the whole curriculum as teachers should not be able to rely on them as the sole means of assessment.

### **4.2.2 Structure**

- 2 Use a variety of activities incorporated in teaching and learning cycles to assess different parts of the curriculum.
- 3 Provide clear guidance on how to carry out tasks and assess pupils including concise instructions, length of the task, any prompts needed, worksheets, marking guidance and examples of pupil responses.

### **4.2.3 Administration and manageability**

- 4 Tasks must have been trialled in realistic situations.
- 5 Allow flexibility for teachers to make adjustments to suit their own context.

- 6 Include ready to use worksheets but provide opportunities for other methods of information collection which teachers can use to level a pupil.

#### **4.2.4 Marking and reporting**

- 7 Provide information in the form of descriptors of performance at each level and pupil exemplars to help teachers interpret results with respect to level in each task. These examples can be provided by pre-testing or trialling the tasks prior to publication.
- 8 Provide guidance on giving feedback to pupils and identifying the next steps. Feedback could include an assessment of what the pupil has answered well, identify areas of weaknesses and next learning steps.

#### **4.2.5 Other considerations**

- 9 For the purpose of professional development, provide examples of tasks and how to assess them so teachers are able to develop their own assessment materials.

## 5. References

- APU (1988). *Science at Age 11. A Review of APU Survey Findings 1980-84*. London: HMSO.
- Bennett, R., Persky, H., Weiss, A. and Jenkins, F. (2008). *Results and Lessons Learned from the NAEP Problem Solving in Technology-Rich Environments Study*. Princeton, NJ: Educational Testing Service [online]. Available: [http://www.iaea2008.cambridgeassessment.org.uk/ca/digitalAssets/180465\\_Bennett\\_R.pdf](http://www.iaea2008.cambridgeassessment.org.uk/ca/digitalAssets/180465_Bennett_R.pdf) [21 October, 2009].
- Bevan, Y., Brighouse, T., Mills, G., Rose, J. and Smith, M. (2009). *Report of the Expert Group on Assessment*. London: DCSF [online]. Available: <http://publications.dcsf.gov.uk/eOrderingDownload/Expert-Group-Report.pdf> [21 October, 2009].
- Crovo, M. and Raizen, S. (2005). 'Development of 2009 NAEP Science Framework and Specifications.' Presentation to the Education Information Management Advisory Consortium (EIMAC), 3 October [online]. Available: [www.ccsso.org/content/pdfs/ScienceFrameworkPresentation.ppt](http://www.ccsso.org/content/pdfs/ScienceFrameworkPresentation.ppt) [21 October, 2009].
- Department for Children, Schools and Families (2009). *Ed Balls' Response to the Expert Group on Assessment* [online]. Available: [http://www.dcsf.gov.uk/pns/DisplayPN.cgi?pn\\_id=2009\\_0090](http://www.dcsf.gov.uk/pns/DisplayPN.cgi?pn_id=2009_0090) [21 October, 2009].
- Foxman, D., Hutchison, D. and Bloomfield, B. (1991). *The APU Experience 1977-1990*. London: School Examinations and Assessment Council.
- Gilmore, A.M. (1999). 'The NEMP Experience:- Professional Development of Teachers through the National Education Monitoring Project.' Paper Presented at the AARE/NZARE Conference, Melbourne, Australia, 29 November-2 December [online]. Available: [www.aare.edu.au/99pap/gil99160.htm](http://www.aare.edu.au/99pap/gil99160.htm) [21 October, 2009].
- Johnson, S. (1989). *National Assessment: The APU Science Approach*. London: HMSO.
- 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.
- Jones, E. and Griffin, H. (2005) *Scientific Minds: a Guide to Assessing Attainment Target One. Key Stage 2*. London: nferNelson.
- Maughan, S (2009, forthcoming). Improving the Acceptability of Teacher Assessment for Accountability Purposes. Some Proposals within an English System. *Cadmo*, XVII, 2, 39-53.

National Assessment Governing Board (2008). *Science Framework for the 2009 National Assessment of Educational Progress* [online]. Available: <http://www.nagb.org/publications/frameworks/science-09.pdf> [21 October, 2009].

National Center for Education Statistics (2009). *NAEP Sample Design* [online]. Available: [http://nces.ed.gov/nationsreportcard/tdw/sample\\_design/](http://nces.ed.gov/nationsreportcard/tdw/sample_design/) [21 October, 2009].

National Foundation for Educational Research (2009). *Submission to Expert Group: Issues to Consider when Developing a National Monitoring System*. Slough: NFER [online]. Available: <http://www.nfer.ac.uk/publications/99901/> [13 February, 2009].

Pyle, K., Jones, E., Williams, C. and Morrison, J. (2009). Investigation of the factors affecting the pre-test effect in National Curriculum science assessment development in England, *Educational Research*, **51**, 2, 269-282.

Qualifications and Curriculum Authority (1997). *English, mathematics, science, Year 4 Assessment Units: Teacher's Guide*. London.

Russell, T. and McGuigan, L. (2003). *Teacher's Guide: Assessing Progress in Science*. London: Qualifications and Curriculum Authority.

Wise, W.L., and C.E. DeMars. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, **10**. 1, 1-17

### **Examples of NAEP hands-on tasks and ICTs**

There is only one hands-on task that is available publicly and this is from 1996, so it is difficult to generalise from this. Pupils are provided with some salt water, some fresh water and a 'mystery' water. By completing the steps of the investigation they must decide whether the mystery water is salt water or fresh water. The steps are very prescriptive but the investigation does tap into misconceptions about measuring and floating.

The framework below gives some examples of possible ICTs.

National Assessment Governing Board (2008). *Science Framework for the 2009 National Assessment of Educational Progress* [online]. Available: <http://www.nagb.org/publications/frameworks/science-09.pdf>

and there are some further examples here:

National Center for Education Statistics (2009). *The Nation's Report Card™: Problem Solving in Technology-Rich Environments. A Report from the NAEP Technology-Based Assessment Project, Research and Development Series* [online]. Available: <http://nces.ed.gov/nationsreportcard/pdf/studies/2007466.pdf>

# 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, QCDA, 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 QCDA or its predecessors. Each of these tests is taken by 600,000 pupils, 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). Members of the Foundation also contribute to the development of the assessments in TIMSS and PIRLS.

NFER was also responsible for the OECD PISA (Programme For International Student Assessment) surveys in England, Wales and Northern Ireland for 2006, and also undertook the 2009 surveys in all four UK countries.