

SATs

the inside story

**The Development of
the First National Assessments
for Seven-year-olds, 1989-1995**



Edited by Marian Sainsbury

nfer

SATS

the inside story

**The Development of
The First National Assessments
for Seven-year-olds, 1989-1995**

Edited by Marian Sainsbury

nfer

Published in October 1996
by the National Foundation for Educational Research,
The Mere, Upton Park, Slough, Berkshire SL1 2DQ

© National Foundation for Educational Research 1996
Registered Charity No. 313392
ISBN 0 7005 1437 6

CONTENTS

Acknowledgements	i
1 Curriculum-Based Assessment and the Search for Authenticity Marian Sainsbury and Steve Sizmur	1
2 Assessing English Marian Sainsbury	16
3 The Development of Assessments of Scientific Investigation John Ashby	32
4 Conflicting Requirements in the Development and Classroom Use of a Task for Assessing Scientific Knowledge Steve Sizmur	42
5 <i>Using and Applying Mathematics: Research into Effective Assessment</i> Eleanore Hargreaves	52
6 An Authentic Test of Mathematics? Eleanore Hargreaves	62
7 Concluding Comments Marian Sainsbury	73
Reference List	78

ACKNOWLEDGEMENTS

There were many other people who worked on the Key Stage 1 development project and are not represented amongst the authors of this book. In particular, thanks are due to Chris Whetton, who directed the project for most of its life, and Graham Ruddock, who led it during its formative years. Other team members were Juliet Burley, Una Christophers, Jen Clarke, Meinir Evans, Maureen Heath, Steve Hopkins, Eleri Jones, Gillian Jones, Keith Mason, Kay McCulloch, John Puncher and Jan Wilson.

The project secretaries were Sandy Williams and Jackie Hill, and invaluable technical support was provided by Mary Hargreaves. The cover design is by Tim Wright, using a photograph taken by John Ashby. The projects to develop Key Stage 1 tasks and tests were sponsored by the School Examinations and Assessment Council and the School Curriculum and Assessment Authority.

1

CURRICULUM-BASED ASSESSMENT AND THE SEARCH FOR AUTHENTICITY

Marian Sainsbury and Steve Sizmur

This book is the story of the early years of National Curriculum assessment for seven-year-old children, told from the distinctive perspective of the SAT¹ development team – those responsible for developing the first ever statutory assessments at Key Stage 1.

It is an account that covers a period of fundamental change in the world of primary education. The introduction of National Curriculum assessment was unprecedented in many respects. It was the first time that any statutory requirements for curriculum and assessment had ever been introduced in this country. The very fact that it was statutory gave it a high profile, so that every development was the object of intense interest. Moreover, it applied to the youngest children, the seven-year-olds, first. The teachers of Key Stage 1 children, unlike their secondary colleagues, were completely unused to any external assessment system. The curriculum itself was introduced without any pilot period, and its assessment requirements were introduced right from the start. Against this background, the work of SAT development was essentially concerned with interpreting and making concrete some of the requirements of the new curriculum. It was a highly innovatory task, but at the same time subject to many difficulties and constraints. Not surprisingly, the six years of the project were years of constant change, both in the assessments we were developing and the curriculum they assessed.

The aim in this book is to explain some of the background that resulted in the form of the SATs that teachers were eventually asked to administer. The project was initiated and steered by Government agencies: at first the School Examinations and Assessment Council (SEAC) and later the School Curriculum and Assessment Authority (SCAA). But at the same time, the development team brought to the work a perspective arising from primary teaching and test development experience. There was a constant attempt to interpret the curriculum faithfully and to develop assessments that matched it well. This

approach to test development was made possible by particular features of the National Curriculum system itself.

The 1988 Education Reform Act was a highly controversial piece of legislation, and its political, social and educational implications have been and will continue to be thoroughly documented elsewhere. From the point of view of this book, however, one feature of the 1988 Act stands out as particularly important. In introducing the National Curriculum, the Act introduced a *curriculum* which was inextricably linked with *assessment*. The curriculum set out programmes of study that were to be taught by teachers, matched by attainment targets setting out what was to be learned by children. It was that same curriculum that was to be assessed. This close linkage between the two had far-reaching implications, and it is some of those implications that this book will set out to explore. The Act set in motion the development of assessments that were required to be *curriculum-based* in a very explicit way.

This point is all the more remarkable when one considers the scope of the National Curriculum. This was no straightforward examination syllabus. On the contrary, the working groups that wrote the programmes of study and attainment targets set out to encapsulate the essential points of good practice in their subjects. In all three core subjects, English, mathematics and science, these essential points included not just knowledge, but skills, understandings and processes.

In each of mathematics and science, a separate attainment target set out the skills and processes that children were to develop. In mathematics, these were the processes of questioning and conjecturing; seeing real-life problems in mathematical terms; investigating within mathematics; looking for patterns and generalisations; and finding ways of communicating about mathematics in words, symbols and diagrams. In science, the processes included raising questions; suggesting ways in which questions could be investigated; controlling variables; making systematic observations; looking for patterns and generalisations; and communicating scientifically in words, diagrams, tables and other media. In both subjects, there was a clear emphasis on the *active* part played by the learner in raising questions and instigating investigations and activities.

Children at Key Stage 1 would be working on the very beginnings of these processes. They should be starting to see the world in

mathematical and scientific terms. The curriculum envisaged the questioning that is so much a feature of young children being encouraged in school, and gradually becoming more systematic and more focused upon matters that are distinctively scientific or mathematical.

The content described in mathematics and science might at first sight seem more straightforward than the processes. But, even here, the curriculum required not just easily identifiable pieces of knowledge, but *understanding*, some of which was fairly complex. Children were expected not just to perform mathematical calculations correctly, but to use their knowledge of number patterns in doing so. In science, children were expected to understand, for example, *how* different forces can act upon an object.

In English, too, there was a broad approach to the subject. Process and content were not set out separately, but both pervaded the five attainment targets of speaking and listening, reading, writing, spelling and handwriting. Right across the subject of English, there was a stress upon *range* of attainments. From the very earliest years in school, talk, reading and writing were to be undertaken for a range of purposes drawn from across the curriculum and from real life.

In reading, it was specified that the point and purpose of the activity – enjoyment of stories, finding out useful information – were to be taught along with the skills necessary for decoding words. For this reason, the reading material was to be intrinsically interesting, enjoyable and of high quality. Similarly, in writing, children were to be taught that writing was a means of communicating information, or of creating stories or poems from the imagination. With each of these real purposes, there was an appropriate form of language activity to be developed. Learning in English was about matching language use to an increasingly wide range of purposes, and, at the same time, helping children to understand explicitly the language choices they were making.

Now, from this brief description of features of the system, the central concern of this book emerges. The assessments were to be curriculum-based. The intention was that assessment results should genuinely give information about how children were doing in the National Curriculum. To be valuable and meaningful, this had to reflect the breadth of the National Curriculum, including its skills, processes and

understandings. Otherwise it could not be taken as an indication of children's progress against the actual curriculum they were required to learn, but only a part of that curriculum, chosen at the convenience of test developers.

What kind of assessment could do this? It is here that the notion of *authenticity* comes to the fore. The intention in setting up a curriculum-based assessment system is to provide assessments that are an authentic reflection of the kinds of work children have to do in following the curriculum.

Typically, the good Key Stage 1 teacher following this curriculum gives children real contexts and purposes within which to investigate and learn. Within these contexts, understanding often emerges in informal conversation with the teacher or with other children. Sometimes, children can convey their understanding in pictures or diagrams. Some of them are becoming adept at expressing their understanding in writing, but this is likely to be a minority in any class. Authenticity must therefore reflect not only the kinds of things to be understood, but the ways in which children of this age are able to show this understanding.

In this sense, the brief given to the SAT developers was an ambitious one. We were to provide tasks that reflected real classroom practice and were meaningful for the children. These assessment tasks were to provide information about the child's progress against a broad and complex curriculum. At the same time as there was this one voice seeking authenticity, however, there were other voices demanding other attributes from National Curriculum assessment, voices which we also had to heed as part of the development brief. These represented a set of demands upon us that were often in tension and sometimes in open conflict. Four main pressures can be identified: curriculum authenticity; reliability; national accountability; and classroom manageability. Each of these presented a distinctive viewpoint, coherent in itself, but not consistent with the other three. In order to represent these conflicts and tensions clearly, the following few pages offer a dramatisation of the debates that affected our work over the life of the project.

THE CAST

- Accountability** – a spokesperson for the people, via government policy.
- Manageability** – a spokesperson for the teachers.
- Authenticity** – a spokesperson for the curriculum and for the children.
- Reliability** – a servant.
- Agency** – a test developer (non-speaking part).
- Narrator.**



ACT 1

- Accountability** Our concern is to improve the education of the pupils of this country. This is to everyone's benefit. We must ensure that standards are raised, and that pupils throughout the country get adequate provision in important curriculum areas. The National Curriculum specifies what pupils should learn, and we must check how schools are doing by assessing pupils against the curriculum.
- Authenticity** Teachers are well placed to assess children's learning, and setting national learning targets will help them to be clearer both about what they are teaching and what children are learning. Teachers can make assessments in the course of their teaching that will give them information on children's future learning needs and help them to aim for those learning targets. They can match assessment to teaching; special tests are not needed.
- Accountability** We need to find out whether standards are improving. That requires comparisons from year to year. It can't be done on the basis of teachers' private assessments, they would be too varied. We need simple benchmarks to which individual children's learning can be compared in the same way in every school. And we need reliable, nation-wide Standard Assessment Tasks at key ages to put them into operation.

- Authenticity** Very well. We shall have SATs. But we must make sure that they represent properly the sort of work children do in school, that teaching to the test will be the same as teaching the National Curriculum.
- The National Curriculum represents good practice in the different subject areas, with a balance of knowledge, understanding and skills. The relationship between process and content must be respected: both must be included so that the integration we want to see in the classroom also features in the tasks. This has implications for the types of task that would be suitable. Modes of presenting the tasks to children and ways for them to respond should allow a faithful reflection of the skills and processes being assessed, and the activities should have a genuine purpose.
- Manageability** The assessments would be more manageable for schools if there were some way of sampling across schools. Each school could carry out assessments of a few areas of the curriculum. Across the country, all the curriculum areas would be covered. That would give us a way to see if standards overall are changing.
- Accountability** But the consumer is at the heart of the education system, and by this we mean pupils and their parents. It's parents that will drive up the performance of schools by sending their children to those giving the best chance of success. To do that effectively, they need accurate information on the attainment of children in each and every school. We must have full coverage, the same coverage, for everyone, and it must overrule teachers' private judgements. Teachers will in any case benefit from more direct and accurate information on how children are doing, so that they can evaluate how successful their teaching has been.
- Authenticity** Then the assessments in every school must reflect accurately the curriculum and its balance. Without that, teachers will not reflect them in their teaching either. We also have to ensure that the ways in which pupils are expected to work are appropriate to the curriculum and to their level of maturity. Because some of the children involved will not even be seven when they are assessed, we must not assume that they can express themselves effectively in writing or in any sort of formal test

situation. We need flexible ways to operate the assessments, using talking, drawing, practical activity or whatever is comprehensible to the children.

Reliability

If we are to have useful information, then we need to make sure we can depend on the results. The assessments must enable children to be allocated a National Curriculum level. We need to make sure that a standard approach is taken across all schools, so that a teacher in Newcastle would award the same level as a teacher in Truro to children of the same ability. Without this, comparisons are meaningless.

The attainment targets are made up of clusters of different attainments. To allocate a level reliably, the assessments must reflect each child's attainment across all the different kinds of performance at the level, represented by the statements of attainment. And because we can't be sure which is the most appropriate level at which to assess a child, we must make sure that he or she has the opportunity to be assessed against the statements of attainment at several levels. For the information to be useful for teachers in planning future learning for children, it is also important to make sure the full range of attainments represented by each statement is included.

**Manageability
& Authenticity**

But we can't assess every child on every aspect of every attainment target.

Manageability

It would take up far too much time.

Authenticity

And we must not subject children with low attainment to tasks that are far beyond their capability, nor should we give children work that is unchallenging. The teacher is the best judge of what tasks the child should undertake.

Reliability

But we can't be sure of that. There's little point in having standard assessments if they can be ignored at will. We need a strategy to ensure that children are consistently given activities at an appropriately challenging level and that where they do well in these, they are then given more demanding tasks.

(The four speakers
turn to Agency.)

All

Help!

ACT 2

Narrator

The year is 1991. Agency returns with a set of SATs in English, mathematics and science. The tasks are designed to reflect quality work in infant classrooms, and feature, for example, assessments of investigations taking place with small groups of children as well as of individual children's reading. They have been developed and trialled in conjunction with schools known for good practice in infant teaching. Solutions have been devised for two main obstacles. Reliability will explain.

Reliability

The descriptions of attainment provided by the statements of attainment are often too vague for teachers to make consistent judgements on the basis of these alone. They have therefore been interpreted further in the context of the specific activities, providing a description of 'evidence of attainment' upon which teachers will base their judgements about individual children. These show the minimum acceptable performance.

The problem of assessing children at the optimal level has been addressed in two ways. In some tasks, it is possible to respond to a single activity at a range of levels, giving 'differentiation by outcome'. That means the same task can be made available to all children, as for example, with writing a story. In other cases (the majority), teachers will enter children first at the level they consider most appropriate. There are then rules which specify whether children should be taken on to the level above or assessed at the level below, depending on the proportion of the questions they get right. This is called 'differentiation by task', and is used, for example, for number work.

Authenticity

We are satisfied that these assessments represent good infant practice as faithfully as possible. Teachers are given flexibility to find ways of making the tasks appropriate to individual children's capabilities and communicative skills. Yet the tasks are capable of providing information that is useful in identifying children's attainments and learning needs, as well as enabling assessments to be made of children's overall level of attainment.

Accountability The curriculum coverage in the attainment targets to be tested is good. We are satisfied that the assessments cover enough to override teachers' own assessments and to set standards for the attainment of pupils that will help teachers focus their teaching and their assessments in the future. We are ready to undertake the first national assessment of all seven-year-old children in England and Wales.



ACT 3

Narrator The first assessment has just taken place in primary schools across the country. A wide range of activities was included. In mathematics, children worked in groups to devise and refine a game, specify its rules and then evaluate it. In science, there was a practical investigation into the disposition of objects to float or sink. In English, children wrote a story about a topic chosen by their teacher and they read individually to their teacher from one of a prescribed list of quality children's picture books.

Manageability The work involved for the class teacher is far too great. Children's education is disrupted for too long. The floating and sinking investigation, for example, took teachers a quite unacceptable amount of time to gather the resources needed and to assess all the children. Scientific and mathematical processes cannot be included in the SAT in future. We need more activities that can be completed by large groups of children together. There should be greater use of worksheets and less use of resources that the teacher has to supply. And then there are all those reading interviews with individual children.

Authenticity But, for the first time ever, we could be sure that scientific and mathematical investigations were taking place in every Year 2 classroom, and that teachers were involved in making assessments of children's work in these areas. The SAT has also required teachers to think about how they can provide for and manage quality work in small groups.

- Manageability** And it has meant that the groups not being assessed were also not being supervised and were given time-filling activities rather than proper curriculum work. This sort of SAT cannot be undertaken by a single teacher unaided.
- Reliability** We can't be satisfied children were being assessed in a consistent way, because the tasks were not sufficiently manageable, because teachers found it difficult to assess children working collaboratively, and because by the time the last children got to do the task, they had found out what to do from those who had already completed it.
- Accountability** What we need to do is reduce the compulsory SAT, and focus it on the basics. Reading, writing, spelling, handwriting and arithmetic should form the core. It is standards in these that concern parents most, and we must make sure we have full information in these areas. For science and the rest of mathematics, we can include a different aspect each year. But no weird experiments over in the corner, please. Wherever we can, we must make it possible to carry out the task in large groups.
- Authenticity** There are two problems with that. These SAT levels are supposed to represent attainment in the National Curriculum, but how can they do that if they don't include the breadth of the curriculum? The related danger is that what gets emphasised in the assessments is what teachers will focus on in their teaching. Skills such as those involved in carrying out investigations and solving problems will get neglected and teacher assessment in these will lack consistency. The 'process' attainment targets must not be lost altogether from the assessments. There should be optional tasks available that will support teachers' own assessments and will convey the standards expected. But once teachers have got used to carrying out the SATs, we must reintroduce these areas into the compulsory tasks.
- Accountability** Perhaps.
- Narrator** Over the next few years, the debate continues. The history of this period is marked by a gradual reduction in the content of the assessments and by increased use of written responses by pupils. The 'process' attainment targets in mathematics and science remain out of the compulsory tasks. One curriculum revision later, there is an organised boycott of SATs by the main teaching unions.

- Manageability** There is still too much for teachers to do, and the revision of the curriculum has made little difference to the amount that must be assessed. But having spelling and comprehension tests is an improvement. These might be a way forward. If only we could dispense with the need for the original SAT approach with all its complexity. We need a radical, full review of the whole system and its purposes.
- Accountability** We must continue to have national testing of the basics. But we should remove the pretence that tests can fulfil the purpose of providing teachers with diagnostic information at the same time as summarising the achievements of a whole key stage. If we change the status of teacher assessment, and report it entirely separately from national tests, that will enable us to optimise the tests to provide a reliable snapshot of the basics.
- Reliability** The use of more standardised tests for reading, spelling and mathematics would make for greater consistency between teachers in awarding levels. Tests should provide for a range of marks at each level, and not focus only on the level thresholds. Getting rid of the statements of attainment would help. They are too restricting. Having more standardised tests would also give all children the same opportunities to show attainment.
- Authenticity** Standardised tests are a worrying feature. It does not follow that giving all children the same question gives them the same chance, because they all have different experiences. And they bear very little relationship to the kinds of activity children do in the classroom. How can they possibly measure attainment in the curriculum? We are concerned that tests will always have a higher profile than teacher assessment, and could start to influence the curriculum experienced by children. However, we do welcome support materials like the optional assessments for scientific investigation. They help to raise the status of teacher assessment and to give useful guidance about approaches and about standards.
- Accountability** We need to take a higher perspective. The introduction of the National Curriculum has been a resounding success. The requirement to carry out the first SATs forced teachers to think about standards of pupils' work and to pay attention to the results of their teaching. The effects

of that will never be undermined. Now we can afford to focus attention more on straightforward, rigorous tests, and leave it to the schools to continue the sterling work they have been doing across the curriculum. And, of course, they will now be inspected far more often, just to make sure.



In this dramatic reconstruction of the tussles between the proponents of authenticity and those with other, insistent, perspectives, a clear shift has emerged in the balance between them over the first few years of the National Curriculum and its assessment. The ambitious programme of authentic, broadly-based assessment was progressively limited in its scope, until 1995 saw the overall thrust of the package as simple, pencil and paper tests wherever possible. There were also, however, cases where a more authentic approach endured.

The chapters of this book will look at the development of assessments over the six-year period of the SAT development project at NFER, in some specific areas of the curriculum. Each one begins with an editorial note to situate its particular focus within the overall structure of the book. Each of the authors will offer an individual perspective on an aspect of the work. In some cases, this is a survey of the direction of development over a period of years. Other chapters choose to focus on particular phases of the development process and draw out the specific issues that emerged over a shorter period of work. In all of them, though, will be found an indication of the notion of authenticity that was guiding us at that time. Some of the inside story of development will be told, so that it becomes clear how the reactions of children, teachers and reviewers fed into the process and led us to modify our initial ideas. The influence of the voices of accountability, manageability and reliability will become clear, and our changing resolutions of the resultant tensions will be described.

This book is aimed mainly at those who have also been involved in the National Curriculum process, but who have other perspectives: teachers, headteachers, advisers and trainers, academics researching the field. We shall not set out to explain in detail the legislation, timetable and curriculum structures that form the background to these accounts. The next three pages, however, set out in tabular form, for reference, the main features of that background.

The Introduction of National Curriculum Assessment at Key Stage 1

Important Dates

Year	The National Curriculum	The SAT Development Project
1988	The Education Reform Act introduces the National Curriculum and its assessment arrangements The Task Group on Assessment and Testing (TGAT) report originates the notion of a 'standard assessment task' National Curriculum Council (NCC) and Schools Examination and Assessment Council (SEAC) set up	
1989	Teachers at Key Stages 1 and 3 start teaching the National Curriculum programmes of study	Three development agencies – NFER, CATS and STAIR – begin to work on the first Key Stage 1 assessments Formal trials of early cross-curricular SATs, summer term
1990		Large-scale pilot by all three agencies Tasks at this stage are still cross-curricular and aimed at assessing everything in the National Curriculum October: NFER appointed as the agency to develop the first national assessments for 1991
1991	Revision of mathematics and science attainment targets SATs criticised for being unmanageable and time consuming	First national assessments, covering nine attainment targets: reading, writing, spelling, handwriting, using and applying mathematics, number, exploration of science and a choice of one further attainment target in both mathematics and science
1992		Second national assessments, covering only seven attainment targets. Practical mathematics and science excluded; more emphasis on pencil and paper approaches Optional SAT pack to support teacher assessment Optional reading and spelling tests October: NFER appointed to develop 1994-6 assessments in mathematics and science but not in English
1993	National boycott of SATs at Key Stages 1 and 3 Sir Ron Dearing appointed to review and streamline the National Curriculum SEAC and NCC replaced by School Curriculum and Assessment Authority (SCAA)	Third national assessments, carried out in only a small number of schools because of the boycott.
1994	Sir Ron Dearing consults on revision of National Curriculum Science excluded from Key Stage 1 SATs	Fourth national assessments, now covering reading, writing, spelling, handwriting and number only Optional number grading tests Optional science pack including exploratory and investigatory science
1995	Introduction of revised National Curriculum Level descriptions replace statements of attainment	Fifth national assessments, covering reading, writing, spelling, handwriting and mathematics Optional Level 2 reading test End of NFER development project
1996		Sixth national assessments, covering reading, writing and mathematics

Structure of the Attainment Targets

1989	1991	1995
ENGLISH		
En1 Speaking and listening En2 Reading En3 Writing En4 Spelling En5 Handwriting		1 Speaking and listening 2 Reading 3 Writing
MATHEMATICS		
Ma1 Using and applying mathematics Ma2 Number Ma3 Number (Operations) Ma4 Number (Estimation) Ma5 Number/Algebra Ma6 Algebra Ma7 Algebra (Graphical representation) Ma8 Measures Ma9 Using and applying mathematics Ma10 Shape and space (Shapes) Ma11 Shape and space (Location) Ma12 Handling data (Collecting and recording) Ma13 Handling data (Representing and interpreting) Ma14 Handling data (Probabilities)	Ma1 Using and applying mathematics Ma2 Number Ma3 Algebra Ma4 Shape and space Ma5 Handling data	1 Using and applying mathematics 2 Number and algebra 3 Shape, space and measures 4 Handling data (not applicable to Key Stage 1)
SCIENCE		
Sc1 Exploration of science Sc2 The variety of life Sc3 Processes of life Sc4 Genetics and evolution Sc5 Human influences on the Earth Sc6 Types and uses of materials Sc7 Making new materials Sc8 Explaining how materials behave Sc9 Earth and atmosphere Sc10 Forces Sc11 Electricity and magnetism Sc12 Information technology and microelectronics Sc13 Energy Sc14 Sound and music Sc15 Using light and electromagnetic radiation Sc16 The Earth in space Sc17 The nature of science	Sc1 Scientific investigation Sc2 Life and living processes Sc3 Materials and their properties Sc4 Physical processes	1 Experimental and investigative science 2 Life processes and living things 3 Materials and their properties 4 Physical processes

Ages, School Years and Key Stages

AGE	SCHOOL YEAR
4-5	Reception
Key Stage 1	
5-6	1
6-7	2
Key Stage 2	
7-8	3
8-9	4
9-10	5
10-11	6
Key Stage 3	
11-12	7
12-13	8
13-14	9
Key Stage 4	
14-15	10
15-16	11

¹ The TGAT report (GB. DES and WO, 1988) originated the term 'standard assessment tasks', which was soon abbreviated to 'SATs'. From 1992 onwards, this abbreviation was no longer used in documents. It remains in current use by teachers and in many articles and other reports. The official term is now 'standard task', or, increasingly, '(standard) tasks and tests'. This book will also make use of the term 'SATs', however, when speaking in an historical context, as that is how published sources referred to them.

2

ASSESSING ENGLISH

Marian Sainsbury

Much of this chapter will focus upon the early years of the development process. It was at the beginning of the development contract, in 1990, that certain fundamental decisions were made about the SAT approaches to speaking and listening, reading and writing. Although there have been changes and adjustments, many of the patterns of assessment established in those early days can still be recognised today. Marian Sainsbury surveys the development work that took place at that time, explores its rationale, and relates it to some fierce public controversies that formed an influential political background. The changes that have been made up to the present can be related to these controversies, as well as to the concerns of manageability and reliability that run through the entire book.

English in the National Curriculum has been, throughout its history, one of the most controversial areas. It is a subject about which feelings run high, and there are deep-seated controversies within it. Even at Key Stage 1, there are aspects of early literacy that seem to generate a great deal of argument and public debate.

At the same period as the National Curriculum was being introduced, there was a bitter controversy about methods of teaching reading and writing, the 'real books versus phonics' debate. Similarly, there were frequent skirmishes on the subject of standard English and its place in the curriculum. Like most such controversies, these rested upon a distortion and caricaturing of what was actually happening in schools, but were powerful influences none the less. The National Curriculum for English was threatened with a review in 1993 in order to reflect some of the views currently in the ascendancy: a review which, in the event, was overtaken by the Dearing review of the entire curriculum.

It is particularly important, therefore, to establish what approach to English the National Curriculum set out, in order to have a firm grasp of the nature of authenticity in English, before going on to consider

speaking, listening, reading and writing in the development of national assessments.

English as a National Curriculum subject, unlike mathematics and science, has been stable in its structure, divided into attainment targets reflecting the four language modes of speaking, listening, reading and writing. Speaking and listening are addressed together; reading constitutes a single attainment target; in writing, there is a recognition of its compositional aspects – *what* is written – on the one hand, and its secretarial aspects – spelling and handwriting – on the other. The four language modes are, in practice, not entirely separate. We read back our own writing, write responses to reading, listen to writing read aloud, and talk about reading and writing. This interdependence is recognised in the programmes of study and the attainment targets, but it is nevertheless generally realistic to identify and assess attainments in each mode separately.

The National Curriculum approach to English has certain features that apply right across speaking, listening, reading and writing. These were set out most clearly in the original Cox proposals (GB. DES and WO, 1989), but can be traced through to the most recent, revised version (GB. DFE and WO, 1995). There is a clear underlying vision of children developing as independent readers and writers, speakers and listeners. The point and purpose of language and literacy use are inseparable from the acquisition of skills and strategies. Real audiences, genuine purposes and a range of stimulating texts are basic requirements. Process and content are addressed together, unlike in mathematics and science. Knowledge about language is important, and takes the form of helping children to understand explicitly the language choices open to them, and the language choices made by other speakers and authors.

Against this broad background, there are specific requirements for each area of English separately. The main part of this chapter will, therefore, consist of separate discussions of speaking and listening, reading and writing, and the nature of authenticity in each will be amplified in the course of each individual discussion.

There is one further factor that distinguishes English from the other core subjects of mathematics and science, in terms of the focus of this book. As the table on page 13 makes clear, the NFER held a contract

as the English development agency from 1989 to the autumn of 1992, so the 1993 national assessments were the last ones developed by the NFER team. After that, development was undertaken for a short while by the East London Assessment Group (ELAG) and then in-house by SCAA. Much of this chapter will therefore focus on the early development process, up to 1992. In fact, some aspects of the materials developed in those early years remained the same through 1994 and 1995. Although, for completeness, this chapter will include some description of development after the end of 1992, this will refer to work not done by the NFER development team.

Speaking and Listening

The discussions and decisions about the role of speaking and listening in the national testing programme took place very early on in the development process, from 1989 to 1990, and their place has remained essentially the same ever since. Nevertheless, the issues that were aired at that time are interesting ones, and it is worth rehearsing them once more, not only for the sake of completeness, but also for an examination of the light they shed on the overall discussion of authenticity that is the theme of this book.

At that earliest phase, the assessment tasks we were developing were cross-curricular in nature. They were based on a theme – *Myself*, for example, or *The World Around Us* – and covered those aspects of English, mathematics and science that related to the theme. Ultimately, this approach was dropped because it proved to be incompatible with the requirement to address all the statements of attainment. A thematic approach worked very well under a loose specification. Then, only those statements of attainment relevant to the theme were addressed, and the resulting activities had very much of the flavour of everyday classroom work. This was the shape of the earliest, 1989, trial SATs. The approach raised considerable difficulties, however, when it came to determining overall results, because only small pieces of evidence were available for any attainment target. By 1990, there was a clear requirement to cover all, or almost all, of the statements of attainment for any attainment target level, and this put a strain on the thematic approach, leading to artificial and cumbersome structures.

The fate of the speaking and listening attainment target was, in some ways, bound up with these early decisions. The early materials included the assessment of speaking and listening as an integral part of other activities, for example, during group practical activities in mathematics and science, during imaginative play connected to the task theme, and during discussion of literature read by the teacher. This structure could be described as authentic in a number of ways. It reflected many of the different types of speaking and listening required by the English curriculum: drama activities; learning through discussion in other curriculum areas; and integrating speaking, listening, and reading in a response to literature. It thus encompassed genuine purposes for children: play, learning or exploring the world of the imagination. The speaking and listening assessments were not, therefore, isolated or artificial activities, but grew out of real purposes for classroom talk. Teachers' responses to the trial tasks and the pilot tasks revealed, however, some attendant problems.

The first was one of manageability. In order to assess speaking and listening, teachers had to set aside time to listen to each of the children as they talked amongst themselves. This in itself, of course, would have provided a valuable experience for teachers. But here it was combined with the need to administer a large number of activities in a short time. The intensity associated with the assessment tasks caused difficulties which would not otherwise have arisen.

A further difficulty with speaking and listening lay in their assessment in a cross-curricular context. If, for example, children were being assessed on their speaking and listening in the context of a science investigation, it was necessary for the teacher to disentangle the language attainments from the science attainments: to attend to different features of the child's words in order to make the different assessments. In assessing science, it was the *content* of the conversation that mattered, for example, whether the child was demonstrating understanding of some scientific concept, or making hypotheses or raising questions. In speaking and listening, on the other hand, the relevance of the content was only one element; others might be the ability to listen to others and to take turns in a discussion, irrespective of whether the science was right or not. Throughout the early development phase, there continued to be disagreement amongst teachers and advisers about how far it was possible for teachers to

make these subtle distinctions. Some argued that it was extremely difficult and that teachers should not be asked to do it. Others held that this view underestimates teachers' professional abilities and that a valuable assessment tool was thereby lost.

A more important dimension to this question, however, is that of the children's own responses in a cross-curricular context. Children who are normally articulate in speaking and listening may be inhibited by a lack of understanding of the science under discussion. Conversely, less articulate children with a good understanding of science may express that understanding better in drawing, writing or practical work than orally. From an assessment point of view, the aim is to give each child the best opportunity to show what he or she can do, and a cross-curricular context can sometimes work against this.

Finally, then, mainly because of concerns about manageability, the 1991 SATs did not include speaking and listening, although a non-statutory task was provided on an optional basis. This situation has continued more or less unchanged ever since, with a requirement to provide a teacher assessment of the attainment target, but no statutory tasks. The 1991 optional task was reprinted the following year, but has not been updated since.

This attainment target provides a particularly acute illustration of the tension between authenticity on the one hand, and all of manageability, accountability and reliability, in different ways, on the other. Compulsory assessments of speaking and listening would have been exceptionally difficult to manage. At the same time, they would be highly variable because of the differing contexts in which they were made, thus causing difficulties for consistency of assessment, which would cast doubt on their dependability. Yet one third of the important subject of English is permanently excluded from the statutory assessments, and any comments on standards in 'English' would be based on literacy only.

It could be argued that reliance on teacher assessment is the only sensible course for speaking and listening, for the best of children's talk – at least with young children – arises spontaneously as a result of interest and involvement in a topic. Teachers have the opportunity to note this talk whenever it occurs, in the course of the whole year's work, across a range of subjects and in a range of groupings. This,

done properly, is surely the most authentic mode of assessment. Yet questions remain about whether it always is done properly. Teachers' practices in assessing speaking and listening have never been exposed to the rigours of standard assessment in the way that mathematical and scientific investigation were in 1991. Many teachers are likely to say that the assessment of speaking and listening 'happens all the time', but the range of those assessments, and the quality of their match to the National Curriculum are still unknown, and probably vary widely.

Reading

The reading task perhaps offers the clearest example of the tensions and conflicts of SAT development, in an area that has always had a high public profile. Our earliest materials included a number of different approaches to assessing reading. One of the early packs, for example, had a single reading booklet with a story reproduced in it; another had flash cards related to the cross-curricular theme, for display and for reading. In 1990, as the pilot phase approached, we concentrated on the development of the type of assessment that seemed, from our early work, to offer the best chance of combining the National Curriculum approach with the necessary degree of standardisation. The work fell into two categories: the choice of text, and the form of the assessment.

It was clear from the programmes of study that children at Key Stage 1 were expected to read a wide range of high-quality texts. The fostering of an understanding of the point and purpose of reading could only take place when the reading-matter itself was stimulating. The standard task needed to reflect this, an important element of authenticity. Rich and stimulating texts cannot be produced to order, and passages specially written to test reading would be unlikely to give rise to the interest and motivation necessary for a valid assessment.

On the other hand, since the results of the assessment were to be taken as evidence of national standards, it was clearly important to show that these results were consistent. But the best children's books are notable precisely for their variety: of text and of illustrations, of voice, tone and vocabulary. To demonstrate this consistency would be a challenge.

From the pilot phase onwards, we decided to base our assessments at all three levels on the kinds of book likely to be found in class book corners, to provide as natural and typical an approach as possible, books such as *Mr. Gumpy's Outing*, *Whatever Next!* or *Greedy Zebra*. For the pilot, a judgement about comparability was made by the research team. For the 1991 SATs, however, it was clear that comparability would need to be clearly demonstrable, and a more sophisticated system of evaluation of texts was put into place. This included a readability measure, the Spache formula (1953; see also Harrison, 1980), but readability measures have considerable shortcomings, especially for picture books. They are mechanical formulae involving, in this case, the proportion of 'unfamiliar' words in the passage and the length of sentences. The Spache formula is unable to give any recognition to supportive page layout, illustrations that give cues to the reader, how repetitive the vocabulary choices are, or how much inference is required to understand the text. Our system, therefore, additionally used ratings for: support from illustrations; line length; and then specific features reflecting the demands at Levels 2 and 3 separately. These ratings were combined into a difficulty score for each text and these scores were kept within a narrow range for all the books at the level.

It is worth spending a little time examining the notion of reliability, or consistency, that underlies this approach. Children – together with their teachers – are offered a choice of books to read for their assessment. It is true that, by using the comparability system described above, the books have been shown to be within a fairly narrow range of difficulty on the five factors taken into account. Nevertheless, it is also true that, in keeping with their quality and originality, the books vary considerably in their actual subject matter, setting, tone, voice and approach. How can this be said to offer a standard assessment? The notion underlying this is that, in this way, children are all offered the opportunity to read something that interests them, with a setting which is reasonably familiar to them. Because children are interested in different things, and are familiar with different physical and cultural environments, then *what* exactly this interesting and familiar setting is, will vary from child to child. And it is in this that the consistency, or fairness, resides. The assessment is standard in the sense that every child gets the chance to read something that is not alien to him or her, and that holds the interest. To take the opposite

approach, and get every child to read the same book, would be consistent in the opposite sense: there would be no difference in what was read, but there would be inconsistency in the familiarity and the interest of that text for individual children. The balance between these two possible approaches also shifted over time. In 1991, there were 27 books on the Level 2 list. This led to concerns about variation, however, and in 1992 the number was reduced to 14. 1993 saw the addition of some titles, bringing the number up to 19, which remained the same until 1996, when only 12 titles appeared.

The second part of our development task was to structure the assessment activity so that children had the opportunity to demonstrate their abilities in reading, as it was defined in the National Curriculum. This went far beyond the word-recognition of traditional reading tests.

At Level 1, the reading activity took the form of a discussion of a book of the child's choice, in the course of which he or she was assessed on interest in books, concept of print, simple comments on content, and recognition of some words or letters. This was really an assessment of emergent reading, and, as such, it did not attract the controversy associated with the higher levels. It was a good example of authentic assessment, differing little from a teacher's normal reading conference with a beginner reader, apart from the structured assessments that were to be made in the course of the discussion. At Level 1, the choice of book was left entirely to the child. This task has essentially stayed the same ever since, with some changes for 1996 to reflect the raising of the difficulty of Level 1 in the revised curriculum. From 1996 onwards, more evidence is required of a child's ability to read – though still with support – than the minimal word and letter recognition in previous years. Correspondingly, there is a list of books to be used for assessment, rather than a completely free choice.

At Level 3, the able readers, too, were originally assessed by means of a reading conference. There was a list of books, and the child read the whole book silently and then talked about the content with the teacher and read a portion of the text aloud. These books were evaluated for comparability using the readability formula and ratings for line length and illustration. The difficulty measure then included the length of the entire story – as children had to read it all – and the degree of scope there was for using inference and deduction in understanding the text.

Here, children were reading silently and beginning to show a deeper understanding of what they had read. The books chosen for the assessment all had scope for this level of analysis. In *Janine and the New Baby*, for example, there was the opportunity to share the feelings of the little girl as her new baby sister is born; in *Ming Lo Moves the Mountain*, the humour of a witty storyline.

The main challenge in devising this assessment was finding a way of ensuring that teachers were given enough guidance to be consistent in their judgements, without restricting children's valid responses unnecessarily. In the pilot, a list of questions was provided for each text, with suggestions for acceptable answers. It was clear that this did not give rise to a coherent and full discussion with the child, so we moved away from that for 1991. Instead, a list was given of points that might indicate the required depth of understanding. Teachers were asked to lead an open-ended discussion with the child, and to assess understanding by following up the child's opening comments.

This task, as at Level 1, exhibited many features of an authentic assessment. In addition to using good-quality books likely to capture children's interest and imagination, it allowed the discussion to range freely so that understanding of the story could be expressed in many different ways, as befits a response to literature.

This reading task was a classic example of the shift in emphasis from authenticity to manageability outlined in Chapter 1. It survived unchanged for three years, 1991 to 1993. From 1994 onwards, however, it was replaced by a written test each year as a consequence of the Dearing review (SCAA was now the development agency). This provided children with a range of texts to read – information as well as stories – and set questions to be answered in writing. It could be taken by all the children at the level, so reducing the time necessary for assessment. It provided a standard set of texts and questions for all pupils. At Level 3, there is a clear expectation that children will be able to read silently and independently in this way. But it is equally clear that this type of assessment, with set questions, a written response mode, and specially-written texts, has lost something of the authenticity of the original task.

It was Level 2 that proved the main battleground over the first five years of the assessment system. Essentially, the question was the

same as at Level 3 – could the reading conference be replaced by a group written assessment, for improved manageability and reliability? Unlike Level 3, this argument has rumbled on for the entire five years.

The characteristics of a Level 2 reader are that he or she should have developed the main strategies for reading – word recognition, phonics, using meaning and grammatical structure – and be able to integrate these strategies and use them as appropriate in gaining meaning from the text. Unlike Level 3, there is not yet an expectation that the child will be able to read silently, or entirely independently of the teacher. This makes a read-aloud assessment the natural approach, as in listening to a child read aloud there is some scope for teachers to identify the strategies the child is actually using.

For this assessment, we developed a running record, to provide a permanent record of the child's reading, on which the assessment could be based. Teachers were asked to note all attempts at words: miscues (that is, words substituted), phonic attempts, omissions and words told, as well as words read correctly. They then had to use this evidence to decide which strategies the child was able to use. For example, a child who tries to sound out a word is showing evidence of using phonic strategies, and one who substitutes a word that makes sense in the context is using grammatical and contextual knowledge. In making the assessment, only words told – which had been read neither independently nor accurately – were 'counted against' the child.

The running record approach was one that had already been recommended, prior to the National Curriculum, by some reading specialists and it had featured in training courses on reading in some authorities. It was not widely used, however, and when it was first piloted, it attracted a mixed response from teachers, trainers and subject specialists. Some critics believed that the method was impossibly complicated for teachers; others that the miscue analysis was so oversimplified as to be completely invalid in this modified form. It received a good deal of attention in the training for the pilot, as at this time it was new to very many teachers. In the event, the pilot reading task proved one of the most popular and successful. Most teachers became reasonably proficient at completing the running record after a few attempts, and the diagnostic value, although not as complete as a full miscue analysis, was recognised by teachers and

advisers alike. The task was therefore taken forward to the 1991 assessments, and has been retained every year since then.

At the same time, however, there have been many criticisms of this task and a series of experiments with other approaches. Some of the most ferocious of the criticisms took place early in the history of the task, in 1990 and 1991. In June 1990, the subject of reading standards came to prominence with the publication, by nine anonymous educational psychologists, of evidence that standards had fallen during the 1980s. In a later pamphlet (1990), Martin Turner, one of the original nine, reiterated these claims and elaborated on his belief that 'real books' approaches to the teaching of reading were to blame for this decline.

The relevance of this debate to the present discussion is not so much about whether standards had or had not fallen, as about the kind of assessment instrument used to measure these standards. The evidence for these claims was based on the practice of a number of local education authorities of administering a standardised reading test to all their pupils at about the age of seven. The use of the same test each year, on large numbers of pupils, allowed a comparison of standards from year to year (Cato and Whetton, 1991). The most common tests in use were group, written, sentence-completion tests. Children were presented with a standard form giving a number of unrelated sentences, each of which contained a space. They were required to select from a number of alternatives the most appropriate word or phrase to fill that space. Only one answer was correct in each case, and the sentences increased in complexity of vocabulary and syntax in order to present increasing difficulty. Whole classes could be tested at the same time, resulting in a relatively quick and manageable form of assessment.

These tests, hitherto the providers of information about standards of reading, differed markedly from the assessment of reading in our tasks. In the latter, we tried to provide an assessment that would reflect both the complexity of the reading process as described in the National Curriculum and normal classroom conditions. Against a background of fierce debate about standards and teaching methods, the SAT assessment was the subject of biting criticism from those who upheld the superiority of standardised tests. Keith Gaines, writing in the *Times Educational Supplement* (1991), described the Level 2 task

as a 'woolly-minded fudge' which was 'clearly biased towards the real books ethic'. Peter Pumfrey and Colin Elliott (1991), in one of a sustained series of criticisms, suggested that the task was 'cruder, more subjective and much less useful than the results of conventional reading tests'.

These criticisms were derived from a view that attached little or no importance to considerations of authenticity, but a great deal to reliability, and 'reliability' defined in a very strict sense. To be reliable, in the eyes of these critics, a test had to use the same passage for every child, and to allow only one correct answer. Clearly the Level 2 reading assessment, with its choice of what to read, its interactive format and its reliance on teacher judgement, did not conform to these criteria.

At about the same time as these technical criticisms were made public, the *Mail on Sunday*, interviewing the then Secretary of State for Education, gave its views on testing reading:

It seems so easy. To Kenneth Clarke it is. But not to the senior educationalists who have wasted millions of pounds of public money trying to decide how to test whether children can read
(*Mail on Sunday*, 16 June 1991)

All this came at the end of the first national administration of the emerging SATs, during which the teaching profession's view was summed up in the national press as twofold. Firstly, teachers had suffered an intolerable workload, and, secondly, they had learned nothing new about their pupils.

The potent combination of political interest and manageability problems led inevitably to a close questioning of the task approach, with Lord Griffiths, then Chairman of SEAC, determined to replace the task at Level 2 with a written test if possible. In the event, a written test was developed for 1992, but was an optional extra. In 1993, there was also a written test alongside the task, but, that year, the teacher boycott led to a very low participation rate in any of the assessments. In 1994 (by which time SCAA had replaced SEAC, and Sir Ron Dearing taken over from Lord Griffiths) the Level 2 task again appeared more or less unchanged. In 1995, the same was the case, but an optional test, *A New Home for Toad*, was produced alongside the task, and was the subject of a large-scale 'technical pilot'. The report

of this exercise (Sizmur *et al.*, 1996a and b) concluded that the task and the test were, in fact, addressing different aspects of reading, and drew the Level 2 boundary in different ways. If forced to choose, teachers would prefer to retain the running record task approach, but most favoured the continued provision of an optional test for the additional information it gave them about some children's reading. Consequently, 1996 saw the retention of the running record task, updated to reflect the holistic approach of the revised curriculum with, again, an optional test alongside. At the time of writing this book, decisions for 1997 were still unclear.

This lengthy chronicle illustrates clearly the way the tensions between authenticity, manageability and reliability exploded into open conflict at times in an area with a high public profile. The considerations brought to bear in the course of these arguments, indeed, went beyond educational and assessment issues, and at times to the heart of political debate.

Writing

By contrast, the writing task attracted relatively few criticisms, either on the grounds of authenticity or manageability. From the start, the task was set up in a similar way to normal classroom practice. Children were to write a story, but teachers had a wide degree of freedom in providing the stimulus and audience for this story. The intention was that, by allowing such a degree of freedom, teachers would themselves ensure the authenticity of the task. The writing curriculum, like other aspects of English, sets out not just that children should learn to write, but that they should develop their understanding of the reasons *why* people write at the same time. That is, they should learn to write in contexts where it is clear that the writing has a purpose, and where they have an audience, a defined *someone* for whom they are writing. In writing a story, children are exploring imaginary events, settings and characters. Teachers were asked to define an audience for the assessment task, so that children could consider what their reader would need to know, as they told the story.

Children with widely varying writing abilities could work on the same task at the same time, and the assessments were made by considering

the overall quality of the resulting piece. Teachers were free to decide whether to run this as a whole class activity, or with smaller groups or individuals, according to their preferences. This flexibility led to a high degree of classroom manageability.

The main difficulty with this approach was that the single piece of story writing was incapable of reflecting the range of writing that children should be attempting. Even at Key Stage 1, they need to be learning about different types of writing for different audiences, and to be considering, for example, how to organise the information in a description or a set of instructions. The story writing task was 'safe'; it did not challenge existing practice in any way, and this was perhaps a reason for its popularity. For 1996, the removal of the constraints of the statements of attainment made it possible for SCAA to develop a task which retained the manageability of the original, but broadened the scope of the types of writing, thus improving its authenticity as a reflection of the entire curriculum. From 1996 on, teachers are invited to consider a range of writing types arising from a variety of stimuli in setting up the writing task. This is itself likely, however, to raise questions of reliability, as it is more straightforward to assess a set of stories against the criteria than to assess stories, instructions, descriptions, lists and other writing types consistently.

Spelling, unlike the compositional aspects of writing, is a more controversial area, and at some stages attracted almost as much attention as reading. In the National Curriculum programmes of study, spelling is envisaged as an active, developmental process in which children make confident attempts at unknown words. They should learn procedures for checking and correcting their work for their audience, alongside a growing vocabulary of words spelled correctly from memory. The approach of the programmes of study was reflected in the original statements of attainment, the basis for the early spelling assessments. Children should exhibit their correct spellings 'in the course of their own writing'. Their attempts at words may be 'recognisable, though not always correct'. The awareness of spelling patterns was credited, in advance of the ability to use those patterns correctly.

The view of good practice embodied in the attainment target required children to take an active role in applying and refining hypotheses about the complex spelling rules and conventions of the English

language. This reduces their dependence on the teacher; it emphasises the importance of audience in deciding upon the standard of correctness to be applied. The original SAT assessment reflected this view. In one piece of writing, children were asked to attempt to spell all the words unaided. They were then asked to use dictionaries to check some of their spellings. This approach was, in fact, unpopular with some teachers at the time. In some classrooms, the children were accustomed to using word books, or resources around the classroom, so that they did not set pen to paper until they were sure of the correct spelling of the word. Whilst being an authentic reflection of the National Curriculum, therefore, it was not as comfortable a match to existing practice as the story-writing itself.

This approach to spelling, however, came under attack in 1991, at the same time as the Level 2 reading task, and for similar reasons. It was argued that the task was unreliable, in that children were not all tested on the same words: by choosing to write only those words they knew how to spell, children could do unfairly well. A straightforward spelling test, it was held, would solve this problem and would, additionally, provide a simple and highly manageable assessment that all the children could take at the same time. As with reading at Level 3, this view prevailed more or less unchallenged, and spelling tests have been used from 1992 onwards. This is another clear example of the movement described in Chapter 1, where considerations of manageability, accountability and reliability have come to the fore, at the cost of some reduction in authenticity.

Conclusion

The original Cox proposals for English in the National Curriculum (GB. DES. and WO 1989) addressed the question of assessment at seven, in advance of the development of any SATs. Here is part of what they said:

A combination of modes of teacher presentation should be used, but pupils' responses should be mainly oral or practical except where the target requires some writing or graphical work by the pupil. The sub-tasks should reflect a range of types of process and of contexts, set in a project which is coherent as a whole and involves group as well as individual activity. (14.18)

... the tasks should be designed to resemble, and build on, normal classroom activities. (14.20)

This recommendation was written at the very outset of National Curriculum English, and shows clearly the overriding concern for authentic assessment that, understandably, informed the conclusions of the working group. They were devising a national system for the first time, and the assessments introduced were to play an important part in supporting the thrust of the curriculum they had formulated. There remain, into 1996, some aspects of the English assessments that fulfil this promise. But there is no hint in the original document of the high-profile debates that were to dominate the progress of English through the first few years of the National Curriculum.

The author of the original proposals, Professor Brian Cox, has chronicled this progress in a book whose title leaves no doubt about the controversial nature of his subject matter, *The Battle for the English Curriculum* (Cox, 1995). Our experiences as SAT developers, which took the form of frequent changes of direction as accountability became more dominant, can be seen as a reflection of the skirmishes, gains and setbacks of this wider battleground.

3

THE DEVELOPMENT OF ASSESSMENTS OF SCIENTIFIC INVESTIGATION

John Ashby

The assessment of young children in science was one of the most innovative features of the introduction of National Curriculum assessment. The curriculum specified that equal weight should be given to pupils' practical skills in exploration and investigation and to their emerging scientific knowledge. This raised the possibility of practical science assessments for seven-year-olds for the first time. In this chapter, John Ashby recounts the challenges of developing these assessments. Like English, this is an area where the conflicts inherent in the development process were apparent at an early stage, and far-reaching decisions were made in 1990 and 1991. Scientific investigation, however, continued to feature throughout the project and the later discussion in this chapter focuses on optional assessment tasks to support teacher assessment, rather than for statutory use.

The first attainment target in science, 'experimental and investigative science', describes the *process* of doing science. Scientific investigation means children working as scientists; children logically and consistently applying intellectual and practical procedures and skills as they investigate physical phenomena in a meaningful way, in order to solve problems or answer (exploratory) questions. Scientific investigations are real tasks, in the sense of being genuine exploration or investigation that is *open-ended*; there is no necessarily preconceived or required solution.

The requirement to develop assessments of scientific process brings with it some very specific and demanding challenges. The assessment of process needs to be carried out on children who are doing a practical investigatory task. The collection of intellectual and practical skills that are necessary to carry out such investigations need to be applied holistically – and, therefore, they need to be assessed holistically. For this reason, assessment tasks for scientific process should consist of entire investigations. At the start of the activity, children should have

the opportunity to identify questions which can be investigated, applying their developing scientific knowledge, and to work out how to test their ideas. Then, they should carry out their tests systematically, observe what happens; measure and record their observations and identify factors which need to be controlled. To round off the investigation, they need to talk about what happened, compare their results with their original ideas and comment on what the reasons might be, using their conclusions to refine their scientific knowledge.

Moreover, these assessments need to be made in a range of contexts – that is, whilst investigating across a range of science. Context has a substantial effect upon children’s performance in process science, and a valid assessment of their process overall must take account of more than one such context.

This chapter briefly chronicles the development of assessment tasks for the assessment of scientific investigation at Key Stage 1 from 1990 and considers issues arising out of that development and some of the implications of the curriculum changes emanating from the Dearing Review (Dearing, 1994a and b).

The Early Tasks

In order to give later developments proper consideration it is necessary to go back to the tasks that we developed for the 1990 trials and to look briefly at how we approached the assessment of science process at that time.

Carrying out assessments through the medium of practical activities was seen as important if the assessments were to lead to valid and authentic judgements of the level of children’s process skills. It was also important that the tasks were set in contexts that were familiar to seven-year-olds, in order for them to be accessible to infant children.

Each of the 1990 pilot SATs contained three practical science activities, set in three different contexts. A child had to complete all three practical activities and the assessment outcomes of each activity were combined to give the child’s overall level in attainment target 1. Requiring each child to do three tasks added both to the authenticity and to the reliability of the assessment. It allowed assessment in more

than one scientific context, and the overall measure that was derived from the combination of a child's assessment outcomes from different practical tasks should be more dependable than one taken solely from the outcomes of one individual practical task.

The total number and nature of the statements of attainment in Sc1 at that time also suggested the need for more than one task. Some statements of attainment were more applicable to particular contexts than others, for example Sc1/3b *identify and describe simple variables that change over time* lent itself well to investigations into the growth of plants, Sc1/2c *use non-standard and standard measures*, needed a context that necessitated the quantification of variables.

The piloting of the SATs in 1990 resulted in three major concerns for the SAT assessment of scientific investigation:

- (1) The manageability of the activities and the workload on teachers and children.
- (2) The problems of group effect whereby teachers had to make assessments of individual children who were working collaboratively in a group.
- (3) The problem of context effect whereby a child's level of achievement in Sc1 showed some variation across tasks set in different contexts. (Whetton *et al*, 1991)

The Floating and Sinking Task

For the 1991 SAT, it was necessary to make an overall reduction in the workload on teachers and children across the core subjects. This was considered to be an imperative but it had ramifications for the other problems with scientific investigation that had been identified in the 1990 trials. The reduction in workload necessitated reducing the number of assessments that teachers were required to make in the SAT. In order to achieve this reduction, only one practical investigatory task could be included.

It could have been argued that the assessment of scientific investigation could not be done through a single practical activity, but had we pursued this option there was the possibility that the assessment of

science process would have been taken out of the SAT assessment altogether – which of course is exactly what did happen in 1992. Yet at that point in time, with the implementation in infant and first schools of the National Curriculum in science being very new, it did not seem desirable to withdraw Sc1 from the SAT. This reasoning appeared to be borne out subsequently to the running of the 1991 SAT when some advisers from different LEAs and also some HMIs reported to us that it was the first time that there had been practical science taking place in all of their schools, as opposed to only in those schools who were ‘pockets of good practice’.

The development of a single practical investigatory task was problematic because it had to meet so many specific requirements and comply with major constraints. The task had to be manageable within the classroom situation that any teacher may find herself in and it had to be set in a context that would be accessible to all seven-year-old children. It also had to promote good curriculum practice by being an authentic reflection of the spirit of science attainment target one.

Apparatus and equipment could be included only if there was a high probability that all schools would already have such pieces or if there was a clear indication from the National Curriculum that schools would have to purchase a particular piece of equipment in order to deliver the curriculum.

The results in terms of children’s levels arising from this assessment had to be secure. Where a disparity occurred between the SAT results and teacher assessment, it was the SAT results that would take precedent. If such were the stakes then it was particularly important to provide every child with the best possible chance of achieving at his or her highest level in the SAT.

As there was only one investigatory science activity then that activity had to be comprehensive enough to accommodate all of the statements of attainment at levels one to three. By making it an extended investigation in order to accommodate all of the necessary statements of attainment we hoped that this would also help to alleviate the context effect. We reasoned that if the investigation was open and extended then there should be sufficient directions for each child to take, in order to follow his or her own particular bent or interest.

We approached the problem of the group effect by putting advice to teachers in the SAT instructions about the possible need to question children further about their individual responses.

The National Assessments of 1991

The 1991 national assessments attracted widespread publicity, and an adverse reaction from many teachers to the difficulties of managing the tasks, especially the practical science. The formal evaluation, however, gave rise to a mixed reaction. The data obtained from the evaluation in 1991 indicated two major issues for us as developers.

- (1) The manageability of an extended investigatory task.
- (2) The problem of the effect of group order whereby groups of children could be inclined to emulate the responses of earlier groups.

The issue of assessment management and classroom organisation was brought sharply into focus by the problems that teachers had with the manageability of the Floating and Sinking Task. Teachers found it difficult to administer this task with a group of children in a classroom whilst being responsible for the management of the rest of the class.

The task was an extended investigation which required the teacher to make observations of each child's contribution towards a group effort. This necessitated the teacher having to give the focal group her undivided attention for an extended period of time. As an aid to manageability, some teachers withdrew children from the rest of the class in order to carry out the task. This strategy was of course dependent upon the availability of another adult to oversee the remainder of the class. There were other teachers who withdrew the focal group from the class as a strategy to overcome the problem of the effect of group order. With the focal group being assessed away from the classroom the rest of the children were not able to observe the responses of the group being assessed, prior to carrying out the task themselves (Whetton *et al.*, 1992).

Both of these problems had been exacerbated by the necessity to have an extended investigation that covered all of the statements of attainment from levels one to three and the need for teachers to

administer the task repeatedly over a restricted period of time. If these requirements could be changed then problems of manageability and of effect of group order could be alleviated. In the event, the difficulties of this task contributed significantly to the 1991 manageability crisis, and the decision not to include any investigative science in future SATs.

The 1991 experience made it clear to us that what was needed was an open investigation that did not require teachers to make too extensive a range of assessments and one that need not be repeatedly administered during a restricted time period to the same class of children. Teachers told us that they had liked the task on its first administration but upon the fifth or sixth administration it had become artificial and stale.

The 1992 Optional Materials

The removal of the assessment of scientific investigation from the statutory assessment tasks in 1992 permitted an easing of the restrictions that had been imposed upon the development of the 1991 activity. There was a similar occurrence in the assessment of process in mathematics. Teachers still had the requirement to finalise their assessments of Sc1 at the end of the key stage but this did not have to be done through the medium of an end of key stage SAT. This was not a precedent as other attainment targets had already been given this same status for the assessment cycle in 1991. For example, 'speaking and listening' had not been SAT-assessed in that year.

This change of status enabled us to provide material in 1992 that contained two practical investigatory tasks set in different contexts; the investigations were called 'Spinning Tops' and 'Growing Plants'. As there was no statutory obligation to use them, teachers had the flexibility of (a) whether to use them at all, (b) in what manner to use them, (c) when to use them and with whom.

The change of status had removed the requirement for the tasks to be administered during a restricted time period, so they could be used at any time during the key stage. It had also removed the necessity to make repeated administrations of any task to the same class of children.

Teachers who were less confident in making assessments of scientific investigation could decide to use them to help in finalising their level judgements. Teachers who were more confident could use them for convenience, there being two ready-made investigations. Teachers were free to use them in different ways, either following the activities closely or using the structure and form of the activities to devise their own.

The format of the assessment materials was identical to the 1991 Floating and Sinking activity, but in structure the 1992 tasks were more open and more truly investigatory. There had been a tension in the Floating and Sinking task between having to have full coverage of all of the statements of attainment whilst giving the task a loose enough structure to enable it to be a true investigation. This problem had been partially alleviated in 1992 by having two tasks that could be flexibly utilised by teachers.

The Role of Material to Support Teacher Assessment of Scientific Investigation

Subsequently to 1992, scientific investigation at the end of Key Stage 1 would not be assessed by a statutory task and because of this we were directed to develop assessment material that would support teachers in carrying out their teacher assessment. In order to carry out this directive, there was a need and the opportunity to clarify the role of assessment support material for teacher assessment of science process.

There is a duality of purpose that support material for teacher assessment must fulfil. This arises from the duality of the purpose that is inherent in teacher assessment. The formative purpose of teacher assessment – for planning teaching and learning – gives it a continuous assessment role, whereas the summative purpose, at the end of the key stage, gives it a different emphasis. One of the main differences between these two roles hinges upon the degree of permanence of the assessment judgements being made by the teacher.

In the formative, continuous assessment role the teacher is making a series of judgements about a child's understanding. Each judgement can be changed by the next set of outcomes that the child exhibits at some time in the future. In the summative role the teacher is making

a collation of judgements which at that point in time is conclusive. If a judgement of a child's performance can be changed by the child's next set of performances then the judgements do not have a *permanent status* but if a judgement is being combined with other judgements in a summation of outcomes then at that point in time each of those judgements does have permanent status. This is, in effect, what is happening when the final level judgement is arrived at by adding together the performances demonstrated by the child. Each of those performances contributes to the final outcome, but in doing so, each is fixed. A problem arises if a child exhibits any atypical performances, which will tend to have a disproportionate effect. This problem can be avoided if a best-fit model of summative assessment is used. In essence, in this approach the teacher takes all of the child's individual outcomes and locates the child within the range of the majority of those different outcomes; atypical outcomes can be disregarded if necessary.

It is a best-fit model of summative assessment that underpins the approach in the booklet of support material for the teacher assessment of scientific investigation that went out into schools for the 1993/94 academic year. This was developed in close collaboration with groups of teachers and the material was quite different to what had gone before. Similar material was concurrently being developed in mathematics.

The 1994 Scientific Investigation booklet contained a range of three investigations set in different contexts. The structure of each of these investigations was minimal, being little more than a framework for each. The aim was to provide teachers with some supporting structure but at the same time avoiding the danger of closing the investigation because of assessment requirements. Included in the framework were indications of the questions that teachers might wish to ask, which in turn gave some indication of the possible directions that different children might take the investigation. A similar approach was taken to mathematics, and is described by Eleanore Hargreaves in Chapter 5.

In order to assist teachers in making assessment judgements about children's outcomes, actual examples of children's responses were included for each of the three investigations. All of the responses were annotated to show how and why particular interpretations and assessment decisions could be made. The inclusion of such examples was seen by our development teachers as being very important and

crucially helpful in supporting teachers in making judgements about their pupils' responses.

The intention was that teachers would use the booklet to support them in both their continuous teacher assessment and in their summative teacher assessment. The material supported teachers in making continuous teacher assessment in that it provided them with exemplifications of children's responses and showed how these related to assessment judgements about those children. It also gave examples of practical activities that teachers could use as assessment vehicles. The material supported summative teacher assessment in that the teacher could consider the children's outcomes from some or all of the practical tasks in the booklet along with the children's outcomes from other practical tasks and arrive at a judgement about the child's typical level of performance. This approach was explained in the introduction to the booklet.

This booklet was well received and during our 1994/95 cycle of evaluations many different teachers made favourable comments.

Development of Support Material 1995

We considered what would be the nature of the support materials that teachers would need when working within the framework of the revised National Curriculum 1995.

The best-fit model of assessment underpinned the changes proposed to the National Curriculum in the Dearing Review. Yet with the replacement of the statements of attainment by level descriptions, teachers needed supporting material that assisted them in locating their teacher assessment judgements within the National Curriculum levels.

The reductions in science content that the Dearing Review proposed had the effect of further limiting the range of suitable contexts in which children's investigatory tasks could be placed. Teachers continued to need material that gave them examples of suitable contexts.

We considered the relationship between a child's level of process skill and his or her level of understanding in the content areas of science. We looked at the use of investigatory tasks as the medium to assess children's knowledge and understanding as well as their level of process skills. We considered the inclusion of this in support materials as one of a range of strategies that teachers could apply in their assessments of children's skill, knowledge and understanding.

Conclusion

Authentic assessment of science process needs to be carried out through the medium of practical investigatory activities. This is time consuming for a teacher to carry out and it calls for a high level of organisation and management skills. Right from the inception of the end of Key Stage 1 National Curriculum assessment, this requirement set up a tension between the manageability of the assessment and its authenticity. In an attempt to resolve this tension the integrity of the practical investigatory task was, at first, compromised in that the task that the children were asked to do was too closed and directed. With the change in status of the assessment of scientific investigation – from being a statutory assessment that all Year 2 teachers had to administer to all Year 2 pupils within a prescribed period, to a non-statutory task – it was possible to build in a high degree of flexibility both within the task administration and in its structure and proposed outcomes. This meant that in the end the change of status resolved the tension between the task's authenticity and its manageability in favour of its authenticity.

4

CONFLICTING REQUIREMENTS IN THE DEVELOPMENT AND CLASSROOM USE OF A TASK FOR ASSESSING SCIENTIFIC KNOWLEDGE

Steve Sizmur

In this second chapter on science, Steve Sizmur considers its other main aspect, scientific knowledge and understanding. The history of this aspect of the subject follows a different course from that of scientific investigation. Tasks for scientific knowledge and understanding remained part of the statutory SATs up to 1993. It is this year that forms the focus for this chapter. Following early experience of using practically based tasks to assess children's knowledge and understanding of science, by 1993 the emphasis was heavily towards making the assessments manageable in the classroom.

This is a story of some of the conflicts that arise in trying to implement standardised external assessments with young children. It shows how those conflicting requirements were faced by those charged with developing the assessments, and ultimately, it shows that it is not possible to resolve the tensions without returning to teachers the faith in their professionalism that the standard tasks were supposed to bypass. The story concerns the assessment of science. Authentic assessment of science in the National Curriculum would include two elements: children's scientific knowledge and understanding; and their skill in conducting scientific investigations. This chapter focuses on scientific knowledge. However, scientific knowledge, as it is portrayed in the National Curriculum, develops as scientific ideas are applied and tested by children against the real world; the processes and products of science may be separated conceptually, but in classroom practice this separation is not so easily maintained.

Statutory SATs for 1993 were shaped by two major requirements. The first was that they should assess as much of the respective attainment targets as possible. This requirement, arising from concerns for accountability in the education system, was related to the need for SAT assessment to override teacher assessment. To do so, they needed to reflect reliably a child's performance across the range of attainment represented by the attainment targets tested. In 1993, this range of attainment was specified by statements of attainment. Hence adequate coverage of these statements was seen as an essential element from the viewpoint of both authenticity and reliability.

This first requirement was in tension with the second: that the assessments be manageable and time-efficient. Experience during the piloting of standard tasks in 1990, and of the first full assessment in 1991, had shown that the need to make assessments of children working on practical tasks in small groups led to a very considerable workload for teachers over the limited period allowed. Hence there was a requirement in subsequent years for as many activities as possible to be capable of being used with large groups of children, up to a whole class. However, not all children in Key Stage 1 are used to working in large groups on the same activity, and when doing so it is difficult to avoid their being influenced by what other children nearby are doing. It is also a problem, under these conditions, to ensure that children understand just how they are intended to interpret the question they are being asked; there can often be areas of ambiguity.

The conflict between these demands became particularly intense in developing a task to assess science attainment target 4 (physical processes). This attainment target featured some quite diverse subject matter: electricity; magnetism; energy sources; energy transfer; forces; light; sound; astronomy. Such a range of content only acquires logical coherence with the development of highly abstract scientific ideas in later years. The need to provide a degree of coherence for the seven-year-old children taking the assessments suggested therefore a further ideal: a single thematic context in which to set the various activities. The thematic 'topic' approach is familiar to teachers of young children, and authentic assessment should recognise this. However, it can lead to artificiality if overdone, and if the theme chosen does not match those being studied in the classroom.

Pupil Sheets

To promote manageability, questions for the 1993 science SAT were presented on pupil worksheets, or *pupil sheets* as they were officially termed. The purpose of a pupil sheet was to give access to the task in a standardised way to as many children as possible. A typical pupil sheet (Figure 1) depicted a situation that should be recognisable to the child, and posed a question or a problem in relation to that situation for the child to address. These questions were often open-ended, allowing children to answer in their own terms. However, as developers, we were aware that presenting a task in a standardised way to all children does not necessarily lead to their construing it in just the way intended. The mere fact that all children are shown the same illustration and asked the same question does not entail that they all see these in the same way. There was also the attendant problem that not all children could be expected to have the reading skills necessary to access the questions in the first place. Hence the degree of standardisation posed a threat to both reliable and authentic assessment of scientific knowledge. Steps needed to be taken to reduce children's difficulties in understanding what was being asked of them.

To address some of the tensions identified here, teachers were given flexibility to:

- ◆ use or discard a unifying theme suggested for the activities;
- ◆ present the tasks to groups of any size, or to individual children;
- ◆ present tasks on pupil sheets, or practically, in a similar or modified context;
- ◆ allow children to read the text of questions themselves, or to present the questions orally;
- ◆ accept children's initial answers, or use further questions to ensure they had answered to the best of their ability. Teachers could also adapt their introductions to ensure full understanding of the questions. Follow-up questioning was recommended in cases where the teacher suspected that children had been influenced by each other.


Figure 1: 1993 Sc4 pupil sheet (shadows)

Sc 4

Name: _____

Sunshine and Shadow

Draw a picture of yourself and your shadow.
Explain how a shadow is made.



Sc 4 Level 2 Pupil Sheet 3 (Sc4/2d) **Sc 4**

This flexibility enabled a balance to be achieved between authenticity and ease of administration. However, inappropriate use of that flexibility could invalidate the assessment. How successful was this compromise? An evaluation of the 1993 assessments was carried out in a sample of 18 schools in England and Wales, and featured both a questionnaire survey involving all the schools and a series of school visits. Nine of the school visits focused on the administration of the 1993 science SAT. Information gathered on these visits and in the questionnaires enabled us to examine how the assessments were used in a range of schools.

Use of the Theme

None of the teachers observed carrying out the assessment had used a linking theme in any way except as a brief introduction to individual activities. No serious attempt was made to link the assessments into a broader classroom context. This was in contrast to previous years, when a majority of teachers had made use of linking themes. Teachers commented that they felt the range of subject matter covered by the task to be greater than they would normally want to cover in a limited period. Some of them accepted in principle that integrating the assessments into a broader range of work was desirable, but pointed out that their teaching 'topics' were fixed long before they had sight of the assessment materials.

The overall impression from the observations made was that teachers wanted to devote as little time to the tasks as possible, and planning and using an overall integrating theme was seen as just one more thing to occupy valuable time; they had given up the attempt to make the assessments relevant. Nevertheless, there was no indication that the teachers were anything other than committed to accurate assessments of their children, and this influenced their use of the other elements of flexibility built into the tasks.

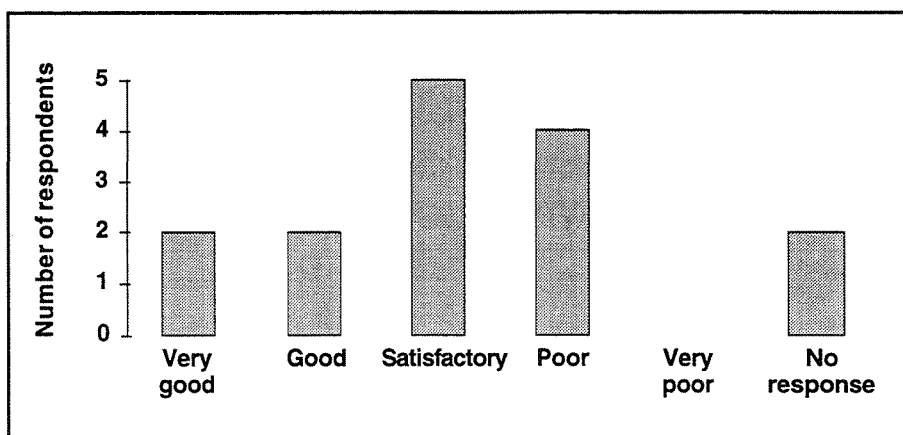
Grouping for the Assessments

Although given the flexibility to administer the assessments with large groups of children, none of the teachers observed chose to work in this way. The larger questionnaire sample revealed that all except one of the teachers had found it necessary to assess some children individually, and one had assessed 24 children in this way.

Discussion with the teachers revealed that they saw small group or individual assessment as the only way to ensure a thorough, fair and accurate assessment. Yet working with small groups meant, for some teachers, leaving the rest of the class to work independently, while others had the benefit of another adult (sometimes another teacher) to assist those not being assessed. So the extent to which teachers could give their attention to the group being assessed varied, and so too did children's opportunities to copy the work of others.

Teachers' views about the manageability of the science tasks were mixed, though one clear message emerged: that the science tasks were less manageable than the two accompanying mathematics standard tasks. Teachers were asked in the questionnaire to rate manageability on a five-point scale from 'very poor' through 'satisfactory' to 'very good'. As many of the teachers rated manageability as good or very good as rated it as poor (Figure 2). None, however, rated manageability as very poor. These ratings contrast with those for the two mathematics tasks, for which none of the teachers gave a rating of less than satisfactory.

Figure 2: Teachers' manageability ratings for the 1993 Sc 4 task

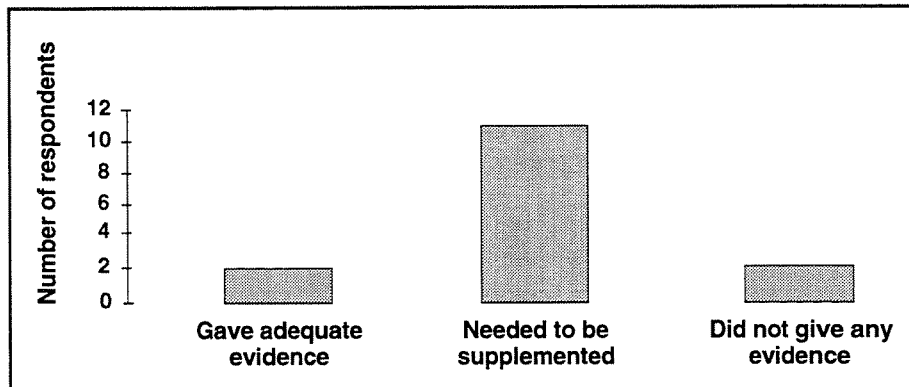


Use of Pupil Sheets

Pupil sheets were used with many of the science activities as a way of promoting access to the tasks for groups of children, and seem to have been accepted as such by most of the teachers. However, it was clear that the sheets in themselves were often an insufficient means of obtaining assessment information. Figure 3 shows teachers' views on the adequacy of pupil sheets for the standard task.

Certain activities were found consistently to require additional follow-up questioning to ensure that children were able to show evidence of attainment, chiefly those activities that involved children in giving an explanation, rather than stating knowledge. This militated against the successful use of these pupil sheets with large groups of children, and at the same time reintroduced a source of potential variation in the assessments. Visits to schools illustrated some aspects of this variation in practice.

Figure 3: Teachers' views on the adequacy of pupil sheets as a means of obtaining evidence



Introduction and Follow-up Questioning

In most cases, the activities were introduced in a straightforward way, and in accordance with the instructions in the materials. Where teachers were seen to deviate substantially from the guidelines, this could invalidate the assessment.

The instructions for administering the pupil sheet on magnetism shown in Figure 4 explicitly stated that children should only be allowed to confirm their predictions about the magnetic properties of the materials shown *after* they had completed the sheet. However, one

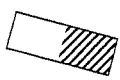

Figure 4: 1993 Sc4 pupil sheet (magnetism)

Sc 4


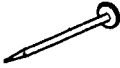



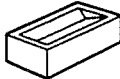




Name: _____

Magnets


Here are some magnets.
Magnets attract things.


Put **a tick** by the things magnets attract.
Put **a cross** by the things magnets will not attract.

orange 	nail 
jumper 	flower 
metal scissors 	brick 
wooden spoon 	paper clip 
drawing pin 	metal spoon 

These two magnets are clinging together.



If one magnet is turned round and they are put together again,
what happens now?



Sc 4 Level 2

Pupil Sheet 1

(Sc4/2a)

Sc 4

teacher ignored this, and let children confirm the accuracy of each of the answers before finalising them. The activity became, therefore, somewhat more authentic as a classroom task, but was no longer an assessment of knowledge. It was a report of an investigation.

A greater degree of variation was to be found in the way follow-up questions were used. One teacher used a very similar 'script' for each small group of children, professing a desire to make the assessments as 'fair' and 'standardised' as possible. During the assessment of children's understanding of magnetic repulsion (Figure 4), two of the children became fixated on the idea that one magnet would spin round, as indeed they could recall having witnessed when they had investigated this for themselves. The teacher's response was to repeat variations of the question 'What else happened when you did it?', and, later, 'Can you think of another word for that', a strategy that did nothing to make the situation real to the children and so help them understand what was being asked. Having discussed the problem at break time, she returned to the children and asked: 'If you were to hold the magnets so that they couldn't spin round, what would happen then?' One of the children then responded that the magnets would push apart.

In other cases, teachers interpreted the freedom to make the task real by additional questioning very widely. A teacher was having difficulty in getting children to explain how their shadow was formed on a sunny day. She considered using the following question, but was unsure of whether this would invalidate the assessment.

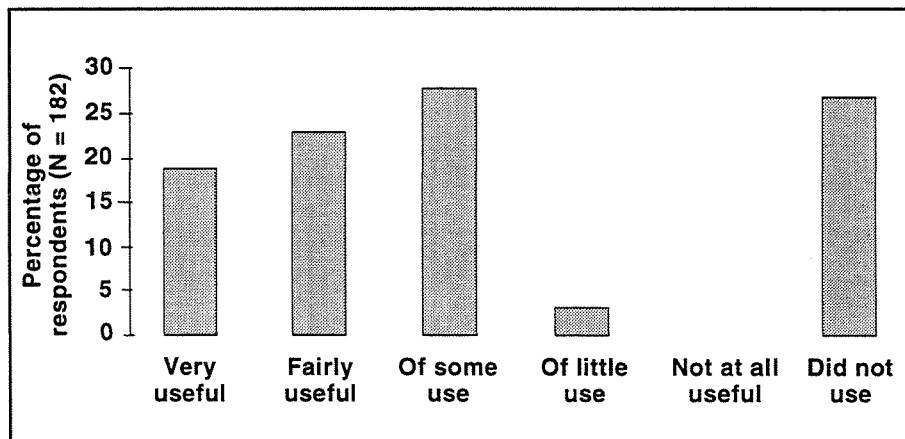
Suppose your Dad's reading the paper, and you stand between him and the light. Would there be a shadow?

Such a question introduces new information and a different context to the assessment. While possibly allowing children more scope to reveal their understanding, this would have resulted in a different activity from that intended. In effect, it would have been teacher assessment, and no longer SAT assessment.

With children's ability to respond thus dependent on the skill and ingenuity of the individual teacher, the capacity for a standard task to overturn the same teacher's judgement was limited. The line between teacher assessment and SAT assessment was far from clear cut in practice. The science SAT could well have contributed to the information teachers had on their children's attainment, and there

were indications that teachers welcomed the availability of the assessment material it provided. To complete this picture, a larger survey carried out in 1994 (Sizmur *et al.*, 1994) revealed that around three quarters of Year 2 teachers had made use of this same science task to support their teacher assessment, and of these the majority found the materials useful (Figure 5). Some of the more useful activities, indeed, turned out to be the very ones that required supplementary questions to obtain the assessment information.

Figure 5: Teachers' ratings of the usefulness of the Sc 4 task in supporting Teacher Assessment in 1994



The 1993 science SAT had taken its place as part of a bank of assessment material drawn on by teachers, as required, to make their own assessments. But it was not the quick and easy-to-use group assessment instrument envisaged in the specification. The range and nature of the subject matter covered (a product of the National Curriculum structure), combined with the quest for a high degree of standardisation, sat uneasily with the flexibility needed to make it an authentic activity in the Year 2 classroom, and militated against its simple integration into the teaching programme within the schools. In subsequent years, the decision to have science assessed at Key Stage 1 by means of teacher assessment alone has removed some of the sources of tension identified here. Following on from this, the NFER team strove to recover some of the ideal, identified by TGAT, that assessment tasks should blend with normal classroom practice. The development of innovative materials to assess scientific investigation, described in John Ashby's chapter, was the first fruit of this new freedom.

5

USING AND APPLYING MATHEMATICS: RESEARCH INTO EFFECTIVE ASSESSMENT

Eleanore Hargreaves

National Curriculum mathematics, like science, has a practical, investigatory element which is presented separately from mathematical content. This, the first of Eleanore Hargreaves's two chapters, addresses the distinctive area of using and applying mathematics. Statutory tasks in this area were discontinued after 1991, but there remained an interest in providing teachers with optional tasks to help them come to grips with the challenges of teacher assessment in using and applying mathematics. After a brief historical overview, the main focus of this chapter is on one particular period in the life of the project where a more open-ended development brief and the involvement of teachers helped us to investigate in depth some of the features of this central aspect of mathematics education.

Using and Applying Mathematics in the National Curriculum

Teachers were required to come to terms with three new curricula over the first five years of the National Curriculum, each with its own particular features. Despite substantial differences among the three curricula, especially their presentation, one aspect of them which has remained relatively constant, was an emphasis on 'using and applying mathematics'.

The ability to use and apply mathematics is set out as a separate attainment target, but, in fact, is intended to permeate and inform all mathematics teaching and learning. In the non-statutory guidance accompanying the original National Curriculum this ability is described as being 'at the heart of mathematics' (NCC, 1989).

There are several strands to the use and application of mathematics which, taken together, are described as mathematical processes. One of these strands is the use of mathematics in a variety of real contexts.

Children need to be able to recognise the mathematical aspects of a problem and select an appropriate mathematical operation; to monitor the results, perform checks and adjust the operation if necessary; and ultimately to complete the solution of the problem.

As they do this, they need to record what they are doing fully and coherently, and to discuss their findings and strategies. Thus, communicating mathematically is a further strand of mathematical process.

There are other mathematical tasks that are not aimed at the solution of a practical problem, but rather that invite exploration within mathematics, to find patterns and generalisations, and for the sheer enjoyment of the exploration. This is a further strand of the ability to use and apply mathematics, and should also be accompanied by a developing ability to record and discuss.

In these mathematical processes, the emphasis is not so much on finding the correct answer quickly. It is more on the development of the ability to tackle problems and questions in mathematical ways, through raising questions, making and testing predictions, and mathematical reasoning. Thus there is an emphasis on open-endedness. It is important to allow scope to explore different approaches, even those that lead nowhere, in coming to grips with a new problem. In mathematical investigations, there is no single right way of doing things, but rather the development of an ability to think mathematically, to use and apply mathematics in a wide variety of contexts.

These features mean that the assessment of 'using and applying mathematics' has particular difficulties inherent in it. Since the emphasis is not on right answers, but rather on appropriate thought-processes, it is exceptionally difficult to provide a standard approach and to give teachers clear guidance on how to assess children's responses. Further, the use of mathematics permeates all the areas of mathematical content, so the actual assessment task will vary quite considerably according to the area of mathematics it is addressing. In order to provide an authentic assessment, children's ability to use and apply mathematics across more than one of these content areas is desirable.

The potential for conflict is already apparent. In order to be authentic, the task must allow more than one right answer, but this detracts from

reliability. A variety of tasks across different areas of mathematics is desirable, but this comes directly into conflict with manageability. In this chapter, I shall outline the research we, as a test development agency, have conducted into the inherent difficulties of teaching and testing 'using and applying mathematics', and I shall discuss some conclusions and outcomes from our research.

Early Assessment of *Using and Applying Mathematics*

The peculiar nature of 'using and applying mathematics' was initially highlighted for us when we became the test agency to develop the first standard assessments of the National Curriculum. As many teachers will remember, the 1991 task required children to design their own game in groups of four, using dice and counters. It was thus quite open-ended in nature, and a variety of examples of acceptable responses was given in order to help teachers make consistent assessments.

In 1991, the mathematics tasks had to be conducted within a limited period and teachers therefore had time only for a single assessment of a child's skills in 'using and applying mathematics'. Yet the task had to provide a faithful reflection of the child's ability actually to use and apply mathematics; and in order that appropriate observations be made, the children had to be assessed in small groups.

Reactions to the standard tasks of 1991 put the future standard assessment of 'using and applying mathematics' into jeopardy. At the time, the main controversy surrounding the standard tasks related to teacher workload and classroom management, both of which many teachers considered unreasonable. The response to the difficulties was to omit the assessment of mathematical processes entirely: in future only the content attainment targets were to be assessed through statutory assessment tasks. Because these attainment targets described content to be learned rather than processes, they were more suitable for written assessments that could be done with a large group. Assessment of process, in contrast, necessarily entailed a more teacher-intensive and time-consuming commitment. The same was true for science. Thus this important aspect of mathematics became a casualty of the conflict between authenticity and manageability.

After that, therefore, the assessment of mathematical process was left up to individual teachers, with sporadic support from local education authorities. In 1992, we did produce non-statutory assessments which were in the same format as the statutory tests but optional. Our large scale questionnaire survey among teachers in 1992 indicated, however, that uptake of these optional materials was limited (fewer than half the teachers in the survey said they used them) (Sainsbury *et al.*, 1992).

Some recent research we have undertaken into systems of teacher assessment (Sizmur *et al.*, 1994), as well as informal research specifically into the assessment of mathematical process, indicated that, in some cases, teachers had not been clear as to exactly what this assessment entailed. For example, some teachers had simply seen it as involving practical work rather than book work, and their teacher assessment levels were based on this misconception. Teachers made comments such as: 'We're doing "using and applying" all the time, so we don't need to make special assessment arrangements'. It seemed that these teachers were focusing on the *practical*, rather than the practical *tasks*; on the *real-life* rather than the real-life *problems*; on mathematics rather than investigation within it.

Research into the Peculiar Nature of *Using and Applying Mathematics*

In the aftermath of the 1991 standard assessments, and also in view of difficulties voiced by teachers in everyday classroom situations, we researched the peculiar nature of mathematical process, in order to find ways in which it might more effectively and easily be assessed. Our analysis implied that the following five characteristics contributed especially to making its assessment different and difficult.

Firstly, despite its separateness, it cannot be assessed in isolation: it must interrelate with other, content attainment targets. But because these others are more concrete, they often become the focus of assessment instead. For example, if you assess process in the context of shape and space, it is easier to concentrate on whether children know their shapes, than on whether they can investigate them.

This tendency is illustrated by the following example, from a classroom trial. As part of an assessment of 'using and applying mathematics', some children were trying to find out which shapes tessellate, and which do not. One child, Ben, made a general statement that 'Shapes with curved edges don't tessellate, but shapes with square corners do'. Ben's teacher was quick to note that many shapes with curved edges do in fact tessellate. In doing so, she was not paying attention to the appropriate mathematical language Ben was using nor to his skills in trying out his own general statement using real examples. Instead she focused on whether he had a correct grasp of the geometric properties of the shapes.

Secondly, as the above example also indicates, assessing children's communication skills and mathematical reasoning necessarily demands intensive teacher attention and observation. Often, assessment of these strands can only be recorded in the form of teachers' observation notes, not as children's own formal work. This makes assessment inherently time-consuming and teacher-intensive.

Thirdly, our research made it clear that children's competence in mathematical communication and mathematical reasoning varies from one context or occasion to another, due both to external and personal changes. Since the skills being assessed depend on a display of each child's initiative to a large degree, the teacher must provide each child with several opportunities to give evidence of his or her full capability. So, not only must the teacher observe intensely, but she must do so in a variety of contexts.

Fourthly, it is not always easy to organise the tasks, problems and investigations necessary. An assessment of process may demand its own different teaching approach and arrangements. It will relate to practical situations and materials. It will also be open-ended enough to allow children to show their skills in predicting, selecting materials, asking questions, solving problems and making investigations. For some teachers, the combination of open-endedness and practicality is complicated to manage.

Fifthly, the necessarily general nature of the statements describing the teaching requirements of 'using and applying mathematics' in the 1989 Order, the 1991 Order and the curriculum for 1995, makes

interpretation difficult unless teachers refer to supporting exemplification. This exemplification is not so important for the other attainment targets since, for them, content is provided rather than process described. The three separate and different, though interrelated, strands each have to be interpreted, assessed and recorded. This complexity is what some teachers failed to acknowledge in the early days of teacher assessment.

Some Conclusions and Outcomes

Out next task was to develop a teachers' booklet of non-statutory assessment tasks in 'using and applying mathematics'. These address directly the complex issues discussed above, and encouraged teachers to focus specifically on this attainment target in its entirety. This time, the non-statutory materials were to be very different from the statutory standard tasks.

Our methodology for developing the approach included the drafting of four activities, which we then trialled ourselves at the earliest stages with groups of Year 2 children. The activities were specifically designed to be integrated into common classroom practices, to make them manageable for teachers and unthreatening for children. Each of the four tasks also had a classroom theme, which included *Fundraising sale*, *Fitting shapes*, *Describing a number* and *Survey*. These provided four differing contexts in which children could show evidence of their skills. Year 1 and Year 2 teachers reviewed and then trialled our early drafts of the activities to ensure their appropriateness.

In order that process could be assessed in the context of the rest of the mathematics curriculum, rather than in isolation, each of the four tasks also drew on one of the other four mathematics attainment targets; for example, the task called *Fitting shapes*, whereby the children tried to find out which shapes do, or do not, tessellate, drew on shape and space. The assessment focus, however, remained securely on process.

Opportunities for displaying the skills embodied in each of the three strands were built into each task. However, expected outcomes were not specified, and therefore it was not anticipated that children would

demonstrate skills from all three strands on each occasion. But there were four tasks, so they were likely to cover all the strands on at least one occasion if they undertook all four tasks at some time during Key Stage 1.

Teacher instructions for progressing through a task, instead of being prescriptive, were in the form of open-ended questions which teachers could select for their group as appropriate. For example, the task called *Fundraising sale* began with the following prompts for teachers:

Ask the children to work out how to raise money for the cause.
The following questions may help:

How many people will come and buy things from the Fundraising sale?

How many items will each person buy?

How much money will each person have?

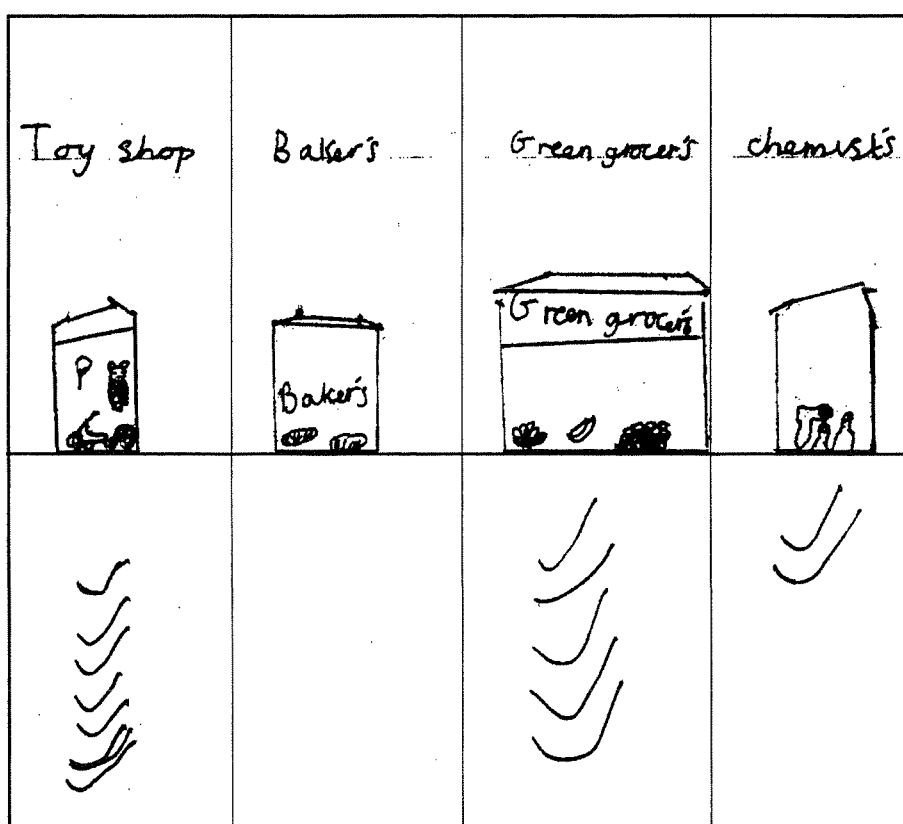
Which items will be most popular? Why?

The next stage of the research exercise involved observing teachers as they conducted the tasks with groups of children in their own classes, and also involved making a substantial collection of children's work as children were assessed. We also collected teachers' notes as ephemeral evidence of a child's performance on a task. This exemplification was especially important in view of the fact that children took routes through a task which were sometimes unexpected, since all the tasks were necessarily open-ended or investigative. The exemplification also addressed ambiguities within the statements.

The examples of children's work and teachers' notes which we assembled formed the basis of a bank of exemplification in the document, which illustrated evidence of attainment at each level and in each strand, for each task. Level judgements were exemplified strictly on the basis of differentiation by outcome. These examples were planned to help teachers interpret the statements, as well as to make their level judgements, and so also to encourage some standardisation of assessment criteria among teachers.

For example, a piece of exemplification from the *Survey* task based in handling data, aimed to help teachers make judgements about the statement in the 1991 order: *respond appropriately to the question: "What would happen if ... ?"*. The exemplification consisted of a child's pictorial tally chart, his teacher's notes, and a commentary, as shown in Figure 1.

Figure 1a. Exemplification: *Respond appropriately to the question: 'What would happen if ... ?'*



This example came out of a large-scale formal trialling of the four tasks, for which 40 Year 2 teachers were randomly selected from across the country. One problem we encountered during this trialling was the necessity of putting a time limit of two weeks on teachers for trialling one of the tasks with a group of children. Ideally there should have been unlimited time. However, the formal trialling exercise

provided us with useful material for refinements of the tasks and boosted our bank of children's work.

A CASE STUDY: SOBAJ - WORKING AT LEVELS 2 AND 3

Sobaj (see Figure 1a)

As well as showing evidence for level 2 in Reasoning, Logic and Proof, Sobaj's teacher's notes about him give other Ma 1 assessment information. He showed evidence of *selecting the materials and the mathematics to use his practical test* (level 2, Applications) since he chose what to survey and how to gather and record his information. The notes indicate that he could *talk about his work using appropriate mathematical language* (level 2, Mathematical Communication), but his teacher might need to talk to him further in order to find sufficient evidence for this. He was, however, able to *present results in a clear and organised way* (level 3, Mathematical Communication).

Teacher's notes (about Sobaj)

Sobaj wanted to find out which shop the children liked to go in best. He chose 4, and asked each member of 2 groups. He said he counted the ticks to find out how many liked each one. He thought the toy shop would be the most popular because children liked toys.

When asked what would happen if there wasn't a toy shop he said more children would be in the greengrocers.

He was alone in suggesting a survey, and he carried it out quite independently.

The booklet was nearly ready for publication and distribution by SCAA when Sir Ron Dearing intercepted progress with his promise of a revised curriculum. The result was that the document could not appear in its current form. However, the findings from the research which led to it may be seen as valuable in their own right, and may be applied to any future curriculum which values process skills in mathematics.

It should be clear from this discussion that the concept underpinning teacher assessment for the revised curriculum, the 'best fit' model, is a concept particularly suited to skills in 'using and applying mathematics'. By the 'best fit' model, a child is assessed on the basis of evidence gathered across time and in different contexts. It may be that, with the emphasis of Key Stage 1 mathematics assessment shifting towards formal written tests, a particular new interest will be stimulated into effective means of assessing process, using the 'best fit' model, since it is an area not assessable by formal written tests. It is now, therefore, even more essential than before that assessment occurs systematically through teacher assessment. Mathematical process appears at the moment to be an area where authenticity is best sought in the context of teacher assessment. To ensure that this is also reliable, teachers are likely to need continuing support in making judgements.

6

AN AUTHENTIC TEST OF MATHEMATICS?

Eleanore Hargreaves

This chapter deals with some of the most recent developments at Key Stage 1, the shaping of the statutory assessments since the Dearing Review of 1994. As our introductory chapter made clear, the emphasis over the six years of the project moved away from the early search for curriculum authenticity and towards a greater concern for manageability and accountability. Sir Ron Dearing's document makes clear that the statutory tasks and tests are summative in purpose and intended to complement teacher assessment. In the mathematics tests of 1995 and 1996, we see the culmination of this process. Most pupils now take a written test covering a range of mathematical content. Eleanore Hargreaves describes the development process for these tests, and also raises the question whether the move away from authenticity in the tests may have gone too far.

Unlike the mathematical processes described in Chapter 5, mathematical *content* has been an enduring feature of the statutory tasks and tests throughout their history. Mathematical content, however, covers a broad range of attainments, including at least number, algebra, handling data, shape, space and measures. Because of this breadth, and the need for classroom manageability, the coverage in the SATs throughout most of their history was limited, emphasising number and in particular number operations, or arithmetic, an important element for national accountability.

Since Sir Ron Dearing's Review of 1993-94, the approach to the assessment of mathematical content at Key Stage 1 has changed radically. This chapter describes the emergence of a mark-based pencil and paper Mathematics Test and raises some questions about the authenticity of the test model.

How the Level 2/3 Mathematics Test Began

In 1995, a significant move was made in mathematics, away from teacher-mediated classroom assessments towards a more formal test which allowed less flexibility of presentation and response. The individual teacher-mediated worksheets which characterised the standard tasks from 1992 up to 1995 were replaced by a test booklet containing 30 mathematics items which all children at Levels 2 or 3 were required to attempt. In effect, teachers had often already used the individual worksheets as if they were pages out of a formal test; but now they had no option. The Teacher's Guide accompanying the test explained that teachers could read all the words in the test questions to the children, but no explanation or probing was allowed: no asking, 'Do you understand what the question is asking you to do?'. Year 2 children had to interpret the words and numbers in the test booklet all by themselves.

This move over to the mark-based test came, ironically perhaps, partially in response to teachers' complaints about the manageability of the earlier, statement of attainment-based tasks. The outcome of the Dearing Review (Dearing, 1994a and b) was a governmental promise of easily administered assessments for the future, and the Mathematics Test was one fulfilment of this promise. The government's own apparent penchant for 'back to the basics short sharp tests' happily fitted into this movement.

The assessment tasks in 1992-1994, and to some extent also in 1991, consisted of pupil sheets setting out written mathematics for children to attempt. This appearance of a written task, however, masked considerable flexibility for teachers. They were obliged to present to the children the mathematical operations set out on the worksheets. But it was open to them to present these operations in any way they liked: orally or practically, rather than in writing. For example, if the worksheet set out shopping sums, the teacher could decide to present these by setting up a class shop and observing the children as they bought and sold items with coins, as long as the number operations were those presented on the worksheet.

In 1994, this approach was subjected to a full-scale evaluation alongside optional written tests (Sainsbury *et al.*, 1994). This showed that teachers in fact made little use of the tasks' flexibility, rarely using oral or practical approaches and rarely presenting them in context. The test booklet approach was viewed favourably from a manageability point of view. In 1995 the Mathematics Task was completely superseded by a single compulsory test based on the 1991 curriculum, and the same model remained for 1996, although in 1996 the test was based on the 1995 curriculum. In this chapter, the 1995 and the 1996 Mathematics Test model will be considered in an attempt to answer the question, 'Is it authentic as an assessment of National Curriculum mathematics?'

What is an Authentic Assessment of Mathematics?

To ask whether an assessment is authentic is to ask whether it is a faithful representation of the mathematical performance required by the curriculum. An authentic assessment reflects the extent and emphases of the field, or construct, being assessed (in this case, National Curriculum mathematics), and allows the child being assessed to display her or his best performance across and within that field.

In summary, an authentic assessment of National Curriculum mathematics should do the following:

- ◆ reflect all aspects of the mathematics curriculum, including 'using and applying mathematics' and 'shape, space and measures'
- ◆ reflect the spirit and emphases of the curriculum, including recognising relationships, developing individualistic mental methods, solving problems through understanding and relating mathematics to purposeful contexts
- ◆ provide for the child to display mathematical skills, knowledge and understanding through a medium or media suited to the child, which could be writing, discussion or practical action
- ◆ reflect the context of the classroom.

How Authentically Does the Mathematics Test Reflect All Aspects of the Mathematics Curriculum?

Standard tasks of 1991 to 1994 had to assess all, or almost all, the statements of the attainment targets they addressed, within the 1989, and then the 1991, Curriculum Order for mathematics. Not so the mark-based test, where a level is achieved by accruing marks rather than by meeting individual criteria. In terms of authenticity, this has both advantages and disadvantages. The thirty items of the test allow for a broader range of mathematics to be included than at any time in the past. Both in 1995 and in 1996, the test included number concepts, patterns and operations, interpretation of data, and two-dimensional shape. The proportions of items in each area reflected the weighting in the curriculum closely. By contrast, the previous standard tasks were restricted to one, or at most two, of the attainment targets.

The attendant disadvantage, however, is that children do not need to show attainment across the entire range in the test to achieve their final level result. Children could be awarded Level 2 in the 1995 test by answering only eight items correctly. Indeed, it was possible to achieve Level 2 without successfully answering a single subtraction question. An error analysis of a sample of 1995 test scripts showed up several children who did exactly that. A mark-based system, then, potentially detracts from the authenticity of the test, in that although more of the curriculum can be sampled, the marks are not attributed separately and children's actual attainments against different aspects of it are not easily evident.

There are particular problems in testing 'shape and space', an area which depends overridingly on the child's practical handling of shapes and interaction with space. Items on two-dimensional shape can be included, but three-dimensional shapes and spatial awareness proved themselves resistant to question writing. In developing the 1996 test, teams of specially trained teachers came to NFER to develop and discuss possible items to include in the test. Each week, a group of them were set the task of thinking up unambiguous items to assess quarter turns and half turns. Each week these items were discussed and 'shredded' and then abandoned during plenary discussion, either for failing to be clear, or for failing to assess exactly

what needed to be assessed. The same happened when one or two items relating to space reached the School Curriculum and Assessment Authority test review panel. No-one seemed to be able to deliver the impossible by producing a written item relating entirely to a physical concept. Three-dimensional shape presented different issues. In order to make assessments of this in a written paper, the three-dimensional shapes had to be depicted by two-dimensional drawings. The recognition and use of a two-dimensional representation of a three-dimensional shape is a Level 6 attainment. Clearly, access to a Level 2/3 test question could not depend upon a higher level attainment in this way.

More serious, however, in terms of overall authenticity, is the omission of the first attainment target, 'using and applying mathematics'. The reasons for this omission were straightforward. A written assessment which forbids teacher intervention and is marked according to a published marking key has little chance of reaching the skills, knowledge and understanding which exemplify the use and application of mathematics. Indeed, the assessment of mathematical process brings with it some very specific problems, as Chapter 5 made clear.

The assessment of the use and application of mathematics was left to teacher assessment alone, while the other attainment targets were at least partially represented in both years' tests. Since the practical application of mathematics, the use of mathematical language and the development of reasoning, logic and proof, as characterised by the first attainment target, are intended to penetrate all other aspects of the curriculum, the authenticity, in pure terms, of a test which systematically omits to assess this section of the curriculum is undoubtedly limited.

Nevertheless, as Chapter 5 makes clear, no statutory assessment since 1991 had included these problematic areas of mathematics, for the reasons of classroom manageability that pervade this entire book. In terms of full reflection of the processes of mathematics, therefore, the tests are found wanting, but no more so than the SATs from 1992 onwards. In terms of breadth of mathematical content, the tests represent a potential gain, but with some qualification.

How Authentically Does the Mathematics Test Reflect the Spirit and Emphases of the Curriculum?

The National Curriculum for mathematics, as contained in the 1995 Order, brings out emphases, in the 'opportunities' sections of the programmes of study, which are less obvious, yet still evident, in its statement of attainment based predecessors of 1991 and 1989. These emphases include recognising relationships, developing individualistic mental methods, solving problems through understanding and relating mathematics to purposeful contexts. An authentic assessment should require children to display their skills in these basic aspects, since these permeate the curriculum against which the children are being assessed. A deliberate effort was made to include assessment of these skills in the 1996 test as the newest curriculum, stressing these skills, was being implemented. Items in this test tended to be more open-ended to encourage children to recognise relationships for themselves. Children were asked, for example, to choose any two numbers that totalled 71, out of a choice of seven numbers, rather than being given a straightforward sum.

A deliberate effort was also made to present items in novel ways so as to challenge children to develop their own methods of solving unfamiliar problems: a bar chart was presented horizontally, rather than vertically as is more usual, so that the child had to interpret the graph in a new orientation and then to deduce that half way between the multiples of ten marked on the axis were multiples of five.

In order to acknowledge a child's understanding of a problem, even if the final answer was incorrect, a system of partial credit was introduced for the 1996 test, in line with practice at Key Stages 2 and 3. For three items in the test, two marks were available: the first for a demonstration of understanding the problem and the second for a process correctly carried out to achieve the answer.

This system, however, is only a token gesture towards authentically assessing the extent of a child's understanding of a problem, which, with children of this age, often requires discussion. The error analysis of children's written responses to the 1995 test threw to light cases where a child's miscomprehension of a question led the child actually to provide a more mathematically sophisticated answer than was

required, and so score no mark. For example, the children were given the numbers 7, 16, 29, 15, 20 and 12 and asked which could be divided exactly by two. A few children divided all of these numbers correctly by two, and wrote the answers, including fractions, in the answer space. Since they did not write just 16, 20 and 12, they were marked wrong. An important element of this question, therefore, was an appropriate understanding of the meaning of 'exactly' in this context. Children who missed this linguistic element gained no credit for their mathematical sophistication. A more flexible assessment approach, where the teacher was allowed to talk through the wording, would have provided a more authentic representation of these children's understanding.

The mathematics curriculum clearly relates the use of mathematics to meaning and purpose beyond that of scoring marks. A test which fails to relate its items to any purpose beyond an assessment of one section of the curriculum lacks authenticity. Even though the aspect of meaning and purpose in mathematics is most strongly emphasised in the first attainment target of the 1991 and 1995 Curriculum Orders, which does not constitute part of the test specification, an attempt was made to make items as meaningful as possible in accordance with the spirit of the curriculum. One method of achieving this was to introduce 'real life' contexts. For example, in 1995, reading a clock was assessed in an item where a picture was provided of a boy putting a cake into an oven, and a sentence began, *Steven put a cake in the oven...* Interviews with children and teachers often indicated, however, that these contexts were hard to relate to, or were simply ignored since understanding them involved extra reading of words.

Data-handling items were found to be useful in providing some purpose to items, since their format could be taken from real life, even if the data itself were fictional. For example, children read off a table of information about some children's heights and weights just as they might have done in their own class.

A third method for injecting meaning into items was to create items which posed their own mathematical challenge. For example, children were asked to use nine triangles to make a bigger triangle. There were various routes to achieving the solution, which therefore involved the child in some problem-solving strategies.

Very many questions, however, had no other purpose than the assessment of one section of the curriculum. For example, children were asked in the 1996 test to write six given numbers in order, starting with the smallest. While this proved to be an efficient means of assessing their grasp of place value, as specified in the National Curriculum, there was no authentic motivation for the operation.

How Authentic is the Mathematics Test as a Medium of Assessment Suited to the Child?

Authenticity, then, is limited by the written mode of testing, since the written mode does not suit the assessment of all areas of the mathematics curriculum, nor facilitate a perception of the child's mental processes and developing understanding. Additionally, the written mode may not suit the child her or himself. Some children may be more hindered than others by assessment through the medium of the written word. It is potentially difficult to disentangle the effects of difficulties in reading from those of a lack of understanding of the mathematics itself: children who are good readers are often good at mathematics as well.

The 1995 evaluation provided a large sample which made it possible to apply sophisticated statistical techniques to this question. A multi-level analysis was carried out, using a sample of over 3,000 seven-year-olds. In this case, the teacher assessment level for mathematics was built into the model, as a measure of children's mathematical ability apart from the test. Even when this had been taken into account, the analysis showed a significant correlation between reading attainment and the final level from the 1995 Mathematics Test (Schagen and Sainsbury, 1996). That is, this analysis seemed to suggest that, for two children with equal mathematics attainment, a better reader would be likely to do better on the Mathematics Test.

Numerous examples could be cited from Mathematics Test scripts where, even without interviewing the child, it is obvious that she or he has answered inappropriately because of a problem with the written mode. During item trials, what had the child read, who drew a dinosaur when asked to draw a pentagon? The child who divided 30

eggs into five neatly illustrated sets of six, but who forgot to write '5' in the answer box, presumably could have counted to five; but by expressing the answer only pictorially the child did not gain the mark.

In sum, there appear to be elements of test-taking technique, or familiarity with the written test format, that hinder access to the Mathematics Test for some children. Where this happens, the test response is not an authentic reflection of the child's mathematical understanding. Although it is clearly important to allow teachers to read the questions aloud to their pupils, this does not provide an entire solution to the problem.

How Authentically Does the Mathematics Test Reflect the Context of the Classroom and Allow the Child to Show Her or His *Best Performance*?

The suitability of the assessment mode, in this case the written mode, relates to the authenticity of the test in the extent to which it reflects classroom practice. Of course, this assumes to some degree that classroom practice is good. Even when it is generally poor, however, at least classroom practice becomes familiar to the children in the class; and some features can be assumed common to nearly all Year 2 classrooms, whether practice is 'good' or not.

The tests were described by one teacher to her pupils as, 'sent to me by the government', and she added, 'I told you they were important'. While many teachers would try to minimise the specialness of the Mathematics Test, the fact that the test itself is alien to the teacher and more so to the children leads to an inauthentic testing situation for all children. The format of the test booklet itself may be alien, the presentation of items may be unfamiliar and the artwork unusual. Whilst attempts were made to make these aspects of the test as accessible as possible, the fact that around 500,000 children took the same test meant that the booklet was bound to be more out of the ordinary for some children than others.

The test booklet was not, then, and could not be, an entirely faithful reflection of normal classroom practice. Perhaps more important was its administration mode, which was out of the ordinary for the vast

majority of classrooms. The main features contributing to this were that, contrary to everyday practice, teachers could not interact with and support children as they worked through the test; they could not ask the children to explain their thinking; children could not discuss their work with each other; and teachers could not select to do the test only with those children whom they considered appropriate.

This last point is an important one. Since the test covered two levels, the final questions were aimed at Level 3, and might be expected to prove difficult or impossible for most of the children taking the test. This placed the children in a situation where they were presented with something they could not do; and rather than being given help in tackling it, as is usual practice, they were told to leave it and go on to the next question. Many teachers in interviews and questionnaire responses commented on the unnaturalness of this situation for their pupils.

In these respects, the testing went against the grain of common infant classroom practice where children work together, discuss with the teacher and use practical situations for mathematics. In particular, infant children do not perform consistently, and revisit a concept at regular intervals in order to revise, extend or reinforce their grasp of the concept. Given a second opportunity to attempt the same challenge, they may use a different method and arrive at a different solution the second time around. The snapshot view of their understanding which is gained by the Mathematics Test may be very limited or simply uncharacteristic, just as one snapshot photograph of a child can depict the child as happy and well turned-out, while another could show the same child as moody and messy. Metaphorically, the teacher is in the position to take a snapshot every day, and so recognise the child's best performances.

Discussion: How Authentic is the Mathematics Test?

Taken alone, the Mathematics Test rates poorly on the scale of authenticity. It does not faithfully and fully reflect the National Curriculum for mathematics at Key Stage 1. Whilst it includes a range of mathematics, and goes some way towards assessing the spirit and emphases of the curriculum, it lacks innate purpose and real-life context. Its authenticity is also limited by the written nature of the

instrument in that it may be unsuitable for or unfamiliar to children in Year 2; its administration mode is not a faithful reflection of common classroom practice; and it gives children only one chance to show their capability in responding to the items included in it.

As one element within an assessment system composed of various assessment procedures, each of which is regarded with equal esteem, the written Mathematics Test could play a beneficial role in the overall authentic assessment of mathematics at Key Stage 1. In a system where the teacher's ongoing observations, discussions, and diagnoses are closely aligned to the National Curriculum and are respected themselves as authentic assessments, the Mathematics Test could provide reinforcing evidence of a child's capabilities; its presentation could aid children in learning to tackle familiar problems through novel means; it could provide diagnostic information about what the child can achieve without help from other children or the teacher; and it could form an integral part of an enjoyable mathematical classroom experience.

This discussion of the Mathematics Test has deliberately been presented in stark terms, taking a strong view of the nature of authenticity in mathematics and in classroom practice. The comments in this chapter serve to highlight in a particularly acute form the most recent progress of the debate between authenticity and the other forces acting upon test development. Taken alone, the current model of the Mathematics Test facilitates ease and consistency of administration and contributes to accountability of results, but is less successful in facilitating the authentic assessment of National Curriculum mathematics. As is clear from both this and the last chapter, it is largely upon teacher assessment that the burden of authentic assessment now falls.

7

CONCLUDING COMMENTS**Marian Sainsbury**

In the chapters of this book, we have described some of our work during the six years of the project and begun to explore some of the issues we have had to come to grips with along the way. The chapters stand alone, each outlining the particular curriculum and assessment issues encountered as part of that work. Other chapters could have been written about other aspects of the work, each casting a slightly different light upon the themes we have identified. It would be premature to offer a set of conclusions, as the system is still developing and the framework for discussing it is not yet established. The following, then, represents some concluding reflections, at the end of this demanding and interesting research project, that seem to emerge from the foregoing chapters, taken overall.

How far do the current national tests and tasks at Key Stage 1 live up to the original ideal of authentic curriculum-based assessments that would reflect and reinforce the programmes of study? The answer is a mixed one. Several of the chapters have chronicled the ways in which more formal written tests have superseded classroom tasks. This move detracts from authenticity in important ways. The aspects of curriculum content that can realistically be assessed by asking a seven-year-old to read questions and write answers are necessarily limited. This narrows the range of the assessment by ruling out any demonstration of deeper kinds of understanding – of literature, for example, or number patterns – that could be assessed in discussion. Also ruled out is the assessment of the processes and skills that permeate the National Curriculum subjects – the only skill to be assessed is the ability to read questions and produce written answers. Under these circumstances, the tests reflect neither content nor process fully, and are in constant danger of superficiality. Furthermore, the unfamiliarity of the test format, and the lack of teacher support may also hinder children's access to the assessment in a significant number of cases.

On the other hand, where classroom tasks have been retained, these dangers are correspondingly less. In reading at Levels 1 and 2, in writing, and in mathematics at Level 1, there are tasks that conform closely to normal classroom practice and would not, therefore, have the disadvantage of unfamiliarity. In reading and mathematics, these are interactive tasks that allow children to demonstrate their understanding in conversation or by practical responses, a more natural medium at this age, and one where the breadth and depth of their understanding is more likely to emerge.

It has to be said that this mixture of tasks and tests appears to attract overall acceptance – though perhaps not enthusiasm – from the teaching profession, in contrast to the outcry of 1991. The provision of supply cover from 1995 onwards has helped teachers to plan and administer the tasks and tests in ways that suit their own organisation, and problems of manageability seem to have been largely solved. The Government, too, seem reasonably happy with this approach to assessing ‘the basics’ for the purposes of national accountability.

The tensions we have encountered in developing the assessments reflect the fact that a national assessment system has many different stakeholders, and that these stakeholders do not all see the requirements in the same way (Daugherty, 1995). It would have been relatively easy for us to devise a set of assessment tasks that gave teachers what they wanted; or that satisfied politicians; or that pleased parents, or subject specialists, or special needs experts, or experts in the technical aspects of test development. The difficulty was to try to do all of these at the same time. This issue can be traced back directly to the original blueprint for SATs, in the TGAT report (GB. DES and WO, 1988), which at the time was hailed as a masterly compromise, satisfying all constituencies. As soon as the model started to be put into practice, however, the tensions within it became very clear.

One path that would still seem to offer further potential is teacher assessment, which has not been the focus of this book, although it has been touched upon from time to time. Since the Dearing Review of 1993-4, the official status of teacher assessment has changed, from a situation where the SAT result was ‘preferred’, to parallel reporting and equal status. This solution is perceived by many as both pragmatic and constructive in terms of educational benefit. It recognises that

teacher assessment defines an entirely different assessment mode, and one which offers great potential advances in authenticity. It is assessment of the curriculum, at the time of teaching that curriculum. It is, however, dependent on the teacher's skill in making these assessments across the whole range of the process and content of the programmes of study, and this is not entirely unproblematic. Our evaluation visits have indicated that, although some teachers deploy an impressive range of observation and questioning techniques, others rely mainly on assessment by means of worksheets that are no more than mini-tests, almost certainly less well thought-out than the national tests.

Since teachers are free to fit teacher assessment into their own classroom routines, its manageability should cause few problems. Here, too, however, the picture is not entirely clear, with one major teaching union strongly in favour of teacher assessment and another regarding it as involving an unacceptable workload.

It is in the area of reliability and accountability, however, that teacher assessment still has an argument to win. Teachers make their assessments in different ways, at different times, of different aspects of the curriculum, using their own interpretations of the standards of performance set out there. All the evaluations of National Curriculum assessment at Key Stage 1 have pointed to the great professional development that takes place in the course of moderation meetings and the preparation of school portfolios. So far, however, there has been a reluctance in the public eye to see teacher assessment as reliable enough to meet the accountability purpose.

It is clear from everything that has been said here, however, that teacher assessment has a vital role to play in providing the broad and authentic assessment that has progressively disappeared from the tasks and tests. John Ashby's account of the optional assessments in science, and Eleanore Hargreaves's of those that were nearly produced for mathematics, have shown the value of this more flexible approach. It has seemed to us throughout the project that the provision of more optional material, offering good classroom activities and guidance on assessment, would achieve the threefold benefit of increasing the reliability of teacher assessment, improving its authenticity for some teachers and, at the same time, raising its status.

For it needs to be said, in conclusion, that the entire enterprise of curriculum-based assessment is a complex and demanding one. Internationally, the introduction of this system has been seen as an ambitious experiment, and it remains the subject of great interest in other countries developing their own national assessment systems. What is at stake is the need to develop an assessment system that will yield accurate information about some very abstract and indefinable educational outcomes. What do we really want of pupils at the age of seven? Almost certainly, we want them to know how to do certain specific things – adding, reading – but we also want them to be interested, curious, thoughtful, able to make sense of new experiences. Even reading and adding, once you scratch the surface, are not simple behaviours at all. Reading means understanding how meaning in language is represented in writing. Adding presupposes an understanding of the number system and how it works. Both of these are highly abstract and complex understandings. Some of the discussion at the beginning of this book showed how the National Curriculum designers did not shrink from reflecting this complexity in the formulation of the curriculum.

So an assessment system that attempts to capture these important features of children's attainment is set a formidable task. The tensions that we have identified throughout this book arise from the intrinsic difficulty of this overall task. One approach is to adopt an assessment mode that seeks to record the full complexity of the child's performance. This is what happens when a teacher records in some detail a child's responses to a classroom activity as part of teacher assessment. It can also be discerned to some extent in the early, practical SATs, where classroom activities were devised and assessments made of children's performances against broad criteria. Half way between these two lie the teacher assessment support materials in science and mathematics: again, these aim at an assessment that captures the complexity of the child's entire response.

An entirely different approach is to simplify the assessment mode, so that the outcome is well defined and easily observable. This is what happens when the content of the programme of study for number, for example, is distilled into a collection of written test items. In this case, the assessment is manageable and more consistent across different

circumstances. But there is a correspondingly greater difficulty in showing that these neat responses actually give evidence about the real underlying understanding that is of interest.

It is for this reason that this book has, in the last analysis, concentrated upon authenticity. Authentic assessments are valuable in that they give direct information about the children's attainments that we really want to know about. Nothing comes between the child's performance and the teacher's interpretation of the understanding shown in that performance. The evidence is direct. But the onus that this places upon teachers should not be underestimated. Their ability to make authentic assessments goes to the very heart of their understanding of the curriculum they are teaching.

REFERENCE LIST

- CATO, V. and WHETTON, C. (1991). *An Enquiry into LEA Evidence on Standards of Reading of Seven Year Old Children*. London: DES.
- COX, B. (1995). *The Battle for the English Curriculum*. London: Hodder & Stoughton.
- DAUGHERTY, R. (1995). *National Curriculum Assessment: a Review of Policy 1987-1994*. London: Falmer Press.
- DEARING, R. (1994a). *The National Curriculum and Its Assessment: an Interim Report*. London: SEAC.
- DEARING, R. (1994b). *The National Curriculum and Its Assessment: Final Report*. London: SEAC.
- GAINES, K. (1991). 'DisSATisfied customer', *Times Educ. Suppl.*, **3901**, 5 April, 29.
- GREAT BRITAIN. DEPARTMENT FOR EDUCATION and WELSH OFFICE (1995). *English in the National Curriculum*. London: HMSO.
- GREAT BRITAIN. DEPARTMENT OF EDUCATION AND SCIENCE and WELSH OFFICE (1988). *National Curriculum Task Group on Assessment and Testing: a Report*. London: DES.
- GREAT BRITAIN. DEPARTMENT OF EDUCATION AND SCIENCE and WELSH OFFICE (1989). *English for Ages 5 to 16: Proposals of the Secretary of State for Education and Science and the Secretary of State for Wales*. York: NCC.
- HARRISON, C. (1980). *Readability in the Classroom*. Cambridge: Cambridge University Press.
- NATIONAL CURRICULUM COUNCIL (1989). *Mathematics: Non-statutory Guidance*. York: NCC.
- PUMFREY, P. and ELLIOTT, C. (1991). 'A house of cards?' *Times Educ. Suppl.*, **3905**, 3 May, 12.
- SAINSBURY, M., ASHBY, J., BURLEY, J., HARGREAVES, E., JONES, E., MCCULLOCH, K. and SIZMUR, S. with SCHAGEN, I. (1994). *Key Stage 1 Mathematics 1994: Technical Pilot Report*. London: SCAA.

- SAINSBURY, M., WHETTON, C., ASHBY, J., SCHAGEN, I. and SIZMUR, S. (1992). *National Curriculum Assessment at Key Stage 1 in the Core Subjects 1992 Evaluation. A Report by the NFER-BGC Consortium*. London: SEAC.
- SCHAGEN, I. and SAINSBURY, M. (1996). 'Multilevel analysis of the key stage 1 National Curriculum assessment data in 1995', *Oxford Review of Education*, **22**, 3.
- SIZMUR, S., CHRISTOPHERS, U. and GALLACHER, S. (1996a). *Technical Pilot of KS1 Optional Reading Comprehension Test for Level 2 for 1995. Final Report*. London: SCAA.
- SIZMUR, S., CHRISTOPHERS, U. and GALLACHER, S. (1996b). 'Where next? Assessment of reading at key stage 1', *British Journal of Curriculum & Assessment*, **6**, 2, 7-11.
- SIZMUR, S., SAINSBURY, M., ASHBY, J. and HARGREAVES, E. with JONES, E. (1994). *Teacher Assessment 1994 Key Stage 1 Evaluation (Mathematics and Science). Final Report*. London: SCAA.
- SPACHE, G. D. (1953). 'A new readability formula for primary grade reading materials', *Elementary School Journal*, **53**, 410-13.
- TURNER, M. (1990). *Sponsored Reading Failure*. Warlingham: Warlingham Park School, Education Unit.
- WHETTON, C., RUDDOCK, G. and HOPKINS, S. (1991). *A Report on the Pilot Study of Standard Assessment Tasks for Key Stage 1, by the NFER/BGC Consortium. Part 1: Main Text and Comparability Studies. Part 2: Appendices*. London: SEAC.
- WHETTON, C., SAINSBURY, M., HOPKINS, S., ASHBY, J., CHRISTOPHERS, U., CLARKE, J., HEATH, M., JONES, G., MASON, K., PUNCHER, J., SCHAGEN, I. and WILSON, J. (1992). *National Curriculum Assessment at Key Stage 1: 1991 Evaluation. A Report on the Working of the Standard Assessment Task by the NFER/BGC Consortium*. London: SEAC.



SATs - the inside story

For anyone involved in primary education, the introduction of Key Stage 1 SATs has been one of the most significant events of the last few years. The floating and sinking task - the booklist - the boycott - all these have become part of teachers' folklore. But what was it like for the people on the inside, the ones who had to write the SATs? For six years, a team of NFER researchers had this task. This book marks the end of the project with a series of reflections on their experiences.

It is a story of constant wrestling with a set of contradictory demands. The SATs had to meet the Government's requirements for reliability and accountability. They had to be a true reflection of the breadth and complexity of the curriculum. Teachers needed to find them manageable in the classroom. And they had to recognise the particular nature of seven-year-olds, the youngest group ever to be formally assessed on a national basis.

This book chronicles how the team dealt with some of these challenges and highlights the fact that there are no absolute solutions to the tensions. It offers some fascinating insights to anyone interested in the impact of the National Curriculum upon Key Stage 1 children and their teachers.

ISBN: 0 7005 1437 6

£9.00