

National Association of Test Directors

We are pleased to present this, the eight volume of published symposia, papers, and surveys of the National Association of Test Directors. The Association considers the promotion of discourses on testing matters from both theoretical and practical perspectives to be essential to its mission. The publication of this document is one important activity that is undertaken to address that goal. These papers were presented at the April, 1992 meeting of the National Council on Measurement in Education in San Francisco, CA. They reflect topics of major interest to members of the measurement and testing communities.

Joe B. Hanson

COLORADO SPRINGS PUBLIC SCHOOLS

Peter Wolmut

MULTNOMAH (OR) EDUCATION SERVICE DISTRICT

Co-Editors

Table of Contents

Symposium I

National Assessment in England and Wales

Introduction

The Assessment System: Purposes and Constraints

Chris Whetton

The Assessment Tasks for Seven Year Olds

Marian Sainsbury

Supporting the Teachers

Steve Hopkins

Changes for 1992

Dorian Bradley

The Development of Tests for 14-Year-Olds

Alan Craig

Advice to US Systems Contemplating Performance Assessment

Chris Whatton

Symposium II

Alternate Assessments in Practice:
Perspectives on Issues & Problems

The Portfolio Reporting Project

Paul LaKahieu JoAnne T. trash

Performance Assessment: Implementation Issues

Maryellen Donahue

Discussion I

Judy Actor

Discussion 11: For a Few Issues More

Peter Wolmut

AUTHORS AND EDITORS

JUDY ARTER Director. Test Center Northwest Regional Zduc..Lab. 101 SW Main St., fto 101 Portland OR 97~204

DORIAN BRADLEY School VxA- . & Assess. Council Newcombe Rout* 45 Notting Rill Cate London W11 US England

MARYELLEN DONAHUE Director, Research & Development Boston Public Schools 26 Court St. Boston MA 02108

JOANNE T ZRZSR Director, Writing & Speaking PLXt2bUrgh Public Schools 341 S. Bellofitld Ave. Pittsburgh PA 15213

ALAN GREIG School Exams. & Assess. Council Newcombe House 45 Notting UL11 Cate Loudon W11 3 JB England

JOZ 3. HANSEN Executive Director Planning. Evaluation, & Research Colorado Springs School District 1115 N. ZI Paso St. Colorado Spring* CO 80903

STEV HOPKINS Bisbop Grossetesto College Lincoln LNI 3DY England

PAUL LEMARIEU Director,,Rsch., Eval., Test Dev. Pittsburgh Public Schools 341 S. Bellefield Ave., Rm. 436 Pittsburgh PA 15213

MARIAN SAINSBURY Nati. Foundation for Educ. Rsch. The Nara Upton Park Slough. Berkshire SL1 2DQ England

CHRIS WHETTON Natl. Foundation for Educ. Rsch. The Nero Upton Park Slough, Berkshire SLI 2DQ England

PETER VOLNUT Director. Measurement/Attendance Multnomah ISD 11611 NE Ainsworth Circle Portland OR 97220

ACKNOWLEDGMENTS

The editors wish to express deepest appreciation to the members of the Board of the National Council on Measurement in Education for their continued support of National Association of Test Directors endeavors. Special thanks go to Nancy Rodgers, Camells, Forster, Sue Aschin, and Charlene Smith for their assistance in producing this Volume.

The views expressed herein are those of the respective authors and not necessarily those of the National Association of Test Directors or its board of directors.

Symposium I

More Authentic Assessment: Theory and Practice

With test directors facing the potential of national assessments, Ernie Bauer (Oakland (MI) Schools) organized this symposium on facets of national performance assessment in England and Wales since 1988. Chris Whatton laid out the fundamentals of the system. Marian Sainsbury then described the assessment tasks created for the seven-year-olds (first group). Steve Hopkins discussed the support provided to teachers in dealing with this new form of assessment. Dorian Bradley presented the changes which were proposed for seven-year-olds in 1992. Alan Greig discussed differences in creating tests for 14-year-olds. And Chris Whetton finished with advice to American systems contemplating the same type of performance assessment.

The Assessment System: Purposes and Constraints

Chris Whatton

National Foundation for Educational Research

Background

A new National Curriculum is gradually being introduced in England and Wales. This is prescribed in the 1998 Education Act, and has four elements: the subjects which are to be taught, Attainment Targets which describe the skills which children should have and what they should know and be able to do, programmes of study which describe the essential matters, skills and processes to be covered, and assessment arrangements for time subjects. The subjects to be included have been laid down as: English, Mathematics and Science which form the core curriculum and foundation subjects: Technology, History, Geography, Music, Art, Physical Education and from the age of 11, a Foreign Language. The manner in which they are to be taught and the materials which are to be used to do this can be determined by teachers themselves using their own methods and approaches.

The model for the structure of the Attainment Targets was determined by a working party, appointed by the government, called the Task Group on Assessment and Testing (TGAT). This was chaired by a science educator, Paul Black. Unusually, and in contradiction to what might be regarded as sound practice, this group was set up to assess assessment arrangements, completely separately from other groups considering the curriculum. The model adopted was to be firmly criterion-referenced, but the dangers of setting minimum competency levels was well recognised, from the North American experience. There was a requirement that the more able pupils should be stretched, showing their maximum attainment. The working group were also attempting to provide a system which would be both formative, allowing teachers to know what to teach children next, giving diagnostic information on strengths and weaknesses and also summative, allowing comparisons between pupils, schools and local education authorities. In fact, the system as a whole was to have five distinct purposes. These were that it should be:

- a. formative - providing information on where a pupil is, enabling teachers to plan
- b. summative - providing overall information on the achievement of pupils
- c. evaluative - providing aggregated information on classes and schools to assess curriculum issues, as well as the functioning of teachers and schools;
- d. informative - providing information to parents about their own children and general information about the whole school;
- e. for Professional development - giving teachers greater sophistication in recording and monitoring so that they can evaluate their own work.

In order to meet these objectives, the Task Group's report (DES, 1987) proposed a scoring system with only ten levels. These would be a continuous scale, independent of age. Level 2 would be, appropriate for the average seven-year-old and an increase of one level would be equivalent to approximately two years of development. This was summarised in a diagram in the report, which illustrated the expected levels which were to be achieved by pupils at certain ages (see Figure 1). Hence an average seven-year-old would be at Level 2, an average 11-year-old at Level 4 and so on. The scale extended up to Level 10, which would have criteria which would stretch the most able 16-year-old. Although the eventual system was to be criterion-referenced this initial norm-referencing would provide the baselines from which the subject working groups could work. Implicit in the model is the greater variance in scores at each age, so that at seven, most children would be at the first three levels, by 16 they might spread over six or seven levels. The working party also confirmed the intention to have assessments made of all children at the age of seven, 11, 14 and 16.

Figure 1: Sequence of Pupil Achievement of Levels Between Ages 7 and 16 [To be included]

In order to demonstrate the progressive nature of the attainment targets, English attainment target 3: Writing can be used as an illustration (see Figure 2). This shows the first three levels for this attainment target. Each level is defined by *statements of attainment* (SoAs). In this case, there is one at level 1, four at level 2 and five at level 3. Some of these SoA form strands which have increasing difficulty (eg. the fun statement in each case), others concentrate on isolated pieces of understanding. There is an inherent assumption that these statements of attainment become steadily more difficult as their levels rise.

Figure 2: Example of an Attainment Target (First Three Levels only)

Attainment target 3: writing

A growing ability to construct and convey meaning in written language matching style to audience and purpose.

Level	Statements of Attainment
Pupils should be able to:	
1	a) use pictures, symbols or isolated letters, words or phrases to communicate meaning.
2	a) produce, independently, pieces of writing, using complete sentences, some of them demarcated with capital letters and full stops or question marks. b) structure sequences of real or imagined events coherently in logical accounts. c) write stories showing an understanding of the rudiments of story structure by establishing an opening, characters and one or more events. d) Produce simple, coherent non-chronological writing.
3	a) produce, independently, pieces of writing; using complete sentences, mainly demarcated with capital letters and full stops or question marks. b) shape chronological writing, beginning to use a wider range of sentence connectives than 'and' and 'then' c) write more complex stories with detail beyond simple events and with a defined ending. d) produce a range of types of non-chronological writing. C) begin to revise and redraft in discussion with the teacher, other adults, or other children in the class, paying attention to meaning and clarity as well as checking for matters such as correct and consistent use of tenses and pronouns.

The attainment targets and the statements of attainment for each subject were first determined by working groups which brought together teachers and others involved in education but also representatives from industry and the wider community. Their proposals were subject to a widespread consultation and, in some cases, considerable public debate before the final version was decided and implemented by the government. One point to note about this process is that those devising the attainment targets were not basing them on empirical data collected on the performance of children but more on their experience and beliefs about what children should be able to do. Hence, they are both targets in the sense of what currently can be achieved and also targets in the sense of aspirations for what ought to become the norm for the future.

Figure 3: Attainment Targets in the National Curriculum

Core Curriculum		
English	En1 Speaking and Listening	
	En 2 Reading	
	En 3 Writing	
	En 4 Spelling	
	En 5 Handwriting	
Mathematics	Ma1 + Ma8 Using and Applying Mathematics	
	Ma 3 Number	
	11 others (knowledge based)	
	(now reduced to 5 attainment targets)	
Science	Sc1 - Exploration of Science	
	16 others (knowledge based)	
	(now reduced to 4 attainment targets)	
Foundation Subjects		
Technology	Design and Technology	4 attainment targets
	Information Technology	1 attainment target
History		3 attainment targets
Geography		5 attainment targets
also		
	Art	
	Music	
	Physical Education	

Finally, in this brief summary of the structure of the National Curriculum, the number of attainment targets in the core curriculum needs mentioning. There were 14 in mathematics, 17 in science, and six in English. Over the first three levels, this led to over 200 statements of attainment. This has now been recognised as too great a number and a revised structure for both mathematics and science has been agreed and implemented. There are now five attainment targets in mathematics and four in science. As a consequence, these are more wide ranging and less coherent but the reduction in numbers has simplified planning, record keeping and reporting.

The total structure of the curriculum is shown in Figure 3. This was the structure which operated in 1991 for the assessments which will be discussed. Hence, in the remainder of the papers of this symposium we will be referring to the old attainment target structure unless the new situation is specifically mentioned.

The Approach to Assessment

As part of the curriculum arrangements, teachers have a duty to maintain records of their children's attainment in all of the attainment targets. This is being achieved through a variety of methods: continuous assessment of the pupils' classwork and homework, classroom tests; informal observation; practical and project work, and school tests. Hence, a large part of the assessment system and the results reported derive from Teacher Assessment. This is to be continuous over the child's time in school and gives the great benefit of a common system closely related to a National Curriculum. Such consistency of education and record keeping for children as they move between schools will in itself provide advantages.

Several of the purposes set out above for the assessment system could be met from teacher assessment alone. In particular the formative function of providing diagnostic information to enable teachers to plan the next stages of a child's education and in providing general information to parents about their own children can happily be accommodated. However, other purposes such as for evaluative information on the class or the school require more formal standardised assessment situations. This is particularly the case if school results are to be published and comparisons between schools and education authorities are to be made. The final purpose, that of professional development, is particularly interesting. It is desirable that teachers should all operate the same standards in their Teacher Assessment but this is difficult to achieve without any element that is compulsory for all teachers.

Hence, an important element of the system as a whole is that there should be some standard assessment or testing at given stages in the education process. In line with the initial recommendations, this is to be at the ages of seven, 11, 14 and 16, which correspond to the times at which most children transfer between schools. Hence, the range of purposes for the national assessment system as a whole, together with the number and nature of the attainment target became important considerations in determining what was to be tested.

The first age group for which National Assessment was carried out was the seven-year-olds. This took place in the summer of 1991. Development work on the style of the assessment had taken place over the previous two years. The formal assessment instruments were to be known, not as tests, but as standard assessment tasks. In order to understand the nature of these standard tasks, it is necessary to be aware of the style of education operating in primary schools. It has been a tenet of British education that at the heart of good primary practice is learning through first hand experience. Young children are thought to learn most effectively when talking, thinking, observing and doing go hand-in-hand. Therefore the activities within the standard tasks were designed to be interesting, practical and relevant to the children. Within primary schools, an activity is rarely limited to

one aspect of learning so the activities presented children with a broad range of coherent experiences, which were aimed at enabling them to demonstrate their knowledge, understanding and skills.

The general philosophy behind the tasks was that, as far as possible, all children were to be given a chance to show what they could do without being limited by the teachers Previous views of them In its purest form, activity was the same for all children. Each child's response to the activity defined the level of attainment The writing task presented an example of this and also how the standard assessment task was actually carried out. Figure 4 shows a page from the teacher's book, demonstrating the form of the administrative procedures for the standard assessment task, which the teacher was required to follow.

Classroom organization varies throughout England and Wales and different types of organisation require different approaches to the assessment. Hence, there were detailed instructions for teachers for each standard task so that they could organise the grouping of children and the balance of curriculum to suit needs of their children within their own approach to teaching. For the purposes of many of the tasks, it was necessary to divide the class into groups of four to six children.

You can see from this that the form of the assessment differs markedly from a written test. The children produce their writing in the way that they normally would within the classroom. There were not test items as such, leading to the award of marks for a number of questions. Instead, the performance required to meet the requirements of a statement of attainment was set out for the teacher, and she had to form a judgement as to whether the child had achieved it. These requirements were referred to as Evidence of Attainment and were an attempt to define the statement of attainment in the particular context of the activity. For writing, the requirements form a hierarchy and the teachers had to decide which was most appropriate to the performance of the child. This style of assessment is referred to as differentiation by outcome and is possible where the statements of attainment form a continuous hierarchy. In other cases, isolated facts or skills make up an attainment target and this must be individually assessed. In this case, we refer to the style of assessment as differentiation by task, and different procedures operated. Marian Sainsbury will describe these in her paper in this seminar.

It was against such a background and in the light of this style of assessment that decisions about what was to be tested had to be taken. Three factors needed to be balanced. These were: the authenticity or validity of the assessments as proper representations of the curriculum; the reliability or consistency of the assessments as fair measures which enable comparisons between school and the manageability of assessments of a large number of attainment targets some of which require individual or small groups to be observed by the teacher.

The first of these, validity or representation of the curriculum was extremely important since the group of teachers had only just begun to teach a newly introduced curriculum It was vital that the correct messages about the relative importance of aspects of the curriculum were given and, moreover, that the style of the

assessment should reflect the spirit of the attainment targets. Hence, for example, in science, the emphasis was to be on children working co-operatively on open-ended practical tasks.

The second aspect, reliability or consistency, was to be achieved through several processes which were 'moderation'. These were the responsibility of local education authorities which consisted of training the teacher, holding agreement trials in which pupil's work was considered, forming networks among schools to jointly consider pupils' work and visiting schools to ensure that the correct standards were being applied. These will be considered in Steve Hopkins' paper. All these were, of course, in addition to the standard nature of the written instructions and Evidence of Attainment given which were performance criteria, defining as closely as possible what was required for children to achieve the Statements of Attainment.

Finally, the total load on the schools, teachers and pupils needed to be considered. At the age of seven, some assessments need to be made on an individual basis, either since the children could not answer a written test or because the nature of the attainment target required it, for example, for reading. For others, small group assessments were required, as for example in science. Some assessments could engage the whole class at one time, for example, writing.

In the light of the development work it was determined that the following should be assessed,

En 2 Reading - an individual activity

En 3 Writing, En 4 Spelling, En 5 Handwriting - a whole-class activity

Ma 1 Using and Applying Mathematics - a small group activity

Ma 3 Number - a small group activity

One further mathematics attainment target

Sc 1 Exploration of Science - a small group activity

One further science attainment target

Hence, in all only nine of the 32 attainment targets were covered in the standard task. Most of these, though, were assessments for small groups. The teachers had six weeks to carry out the assessments, arranging their own timetable to cover them in the order they wished. The estimated time to complete these was 30 hours, excluding the testing of reading. Teachers were expected to have been keeping records and assessing all the attainment targets, so that for those not covered in the standard assessment task the Teacher Assessment stood alone.

Any evaluation of the successes and failures of the first year of the National Assessment has to take into account the five purposes of the system as a whole and how each of these were met. Inevitably, reactions of different participants reflected their own judgements as to the most important of the purposes. Such views of

the priority of purpose also determined the attitudes to the balance struck between curriculum validity, reliability and manageability. These reactions and the response to them will be described in the papers which follow.

The Assessment Tasks for Seven-Year-Olds

Marian Sainsbury

National Foundation for Educational Research

Introduction

The opening paper by Chris Whetton has explained the principles underlying the form and style of the standard assessment tasks. Essentially, these tasks were to reflect normal classroom practice as far as possible. This requirement enhanced the validity of the assessments, both as reflections of the curriculum and as indicators of young children's attainment.

The assessments were directly related to the programmes of study of the National Curriculum and were designed to reflect that curriculum by including practical and problem-solving skills as well as knowledge and understanding. In assessing children as young as seven years, it was recognised that they needed flexibility of response mode, so that they could express their understanding orally or practically if they were unable to read and write. These requirements gave rise to a highly unusual form of assessment instrument, consisting not of test papers, but of a set of instructions for teachers, combined with a recording booklet specifying acceptable responses.

The tasks themselves varied according to the nature of the skills, knowledge or understanding to be assessed. This paper will illustrate this variety by describing in more detail the approach to the assessment of number, of exploration of science and of reading in 1991. The tasks were evaluated by means of a questionnaire survey and a series of case studies, and evidence from this evaluation will be used to describe some of the reactions of teachers and children to the tasks.

Assessing Number

Mathematics attainment target 3 sets out the number skills required of young children at the first three levels of the National Curriculum: at level 1, simple calculation to ten; at level 2, more advanced addition and subtraction; at level 3, addition, subtraction, multiplication and division, and the use of calculators. Figure 1 sets out the statements of attainment addressed by the tasks at each of these levels.

This attainment target is an example where it is not practicable to devise a task allowing differentiation by outcome -that is, where all the children attempt the same work and an assessment is made of their differing responses. Here, it would clearly be wasteful of time to ask an able child to carry out very simple calculations, and demotivating to ask a less advanced child to do the demanding work required at level 3. For this task, therefore, teachers were asked to decide upon an 'entry level', using their teacher assessments, and to explore upwards and downwards as necessary after that. For this attainment target, therefore, differentiation by task was employed.

Figure I

Ma 3: Number

Level 1

Ma 3/1a

*Add or subtract, using objects where the numbers are no greater than 10***Level 2**

Ma 3/2a

Know and use addition and subtraction facts up to 10

Ma 3/2c

*Solve whole number problems involving addition and subtraction, including money***Level 3**

Ma 3/3a

Know and use addition and subtraction number facts to 20 (including zero)

Ma 3/3b

Solve problems involving multiplication and division of whole numbers or money, using a calculator where necessary

Ma 3/3c

Know and use multiplication facts up to 5x5, and all those in 2, 5 and 10 multiplication tables

There are three statements in the list, 2a, 3a and 3c, that require children to 'know and use' number facts in addition, subtraction, multiplication or division. This contrasts with the ability to perform calculations, in that it requires the instant recall of the answer. In the standard task, this was assessed by asking children to play a dice game -sometimes with specially prepared dice where they had to throw two dice, perform the correct operation on the two numbers thrown, and give the answer as quickly as they could.

The instructions to the teachers said:

It is important that you assess each child's ability to add and subtract by using recall of number facts only, not by counting or computation. These methods can only be properly distinguished by observation so you will need to watch the children carefully.

This requirement to observe children closely and to make a judgement about how they had arrived at their correct answers placed a considerable burden upon teachers. Firstly, they had to organise their classes so that they were able to give concentrated attention to the children being assessed. Then, they had to make a judgement about children's means of arriving at an answer.

Case study observations showed a wide variety of practice. Children of this age - as is probably the case for most adults - have evolved a wide range of strategies for giving such answers. Some of them genuinely involve recall. Typically, children were observed to give answers to simple sums such as $4 + 4$ instantly, from memory, without any visible calculation. But more difficult number facts, $8 + 5$ for example, gave rise to counting on fingers, counting on mentally, and other strategies that can only be guessed at, many of which provided very quick answers. Teachers were bemused as they attempted to distinguish acceptable from unacceptable

responses. Some made their pupils sit on their hands, to avoid finger-counting. For some, a response had to be instant; others allowed a few seconds.

There was a further problem, in that the dice game device did not always give rise to assessments of the same level of difficulty. Any child repeatedly throwing $2 + 2$ had a clear advantage. It was open to teachers to give further throws until they were satisfied that the task was comparable, but some disquiet remained about consistency of assessment.

By contrast to this, the assessments of the statements that allowed computation were relatively simple to administer. Worksheets were provided with shopping calculations set out on them. Some of these are reproduced in Figure 2. [to be included] Many teachers chose to approach this as a formal pencil-and-paper exercise. But those teachers who wished to adopt a more informal approach, matching their usual practice, had the option of making these assessments orally, by labelling items in the 'class shop' with the same prices and asking children to perform the calculations in the course of imaginative play.

Some of the assessments of simple number operations, therefore, came close to looking like fairly conventional tests, albeit not carried out under formal testing conditions. The main problems in this area arose where the National Curriculum required that teachers should assess not just the correctness of the answer but also the mental processes that gave rise to it: Dorian Bradley's paper will describe the changes that were made to the 1992 tasks in an attempt to address these problems.

Assessing Exploration of Science

The science attainment targets in the National Curriculum, as applied in 1991, are 17 in number. Sixteen of these cover discrete areas of content, but attainment target 1 differs from these in that it gathers together the process skills that are to be applied across all content areas. As such, exploration of science poses particular challenges for assessment.

In an attempt to meet these challenges, the 1991 standard assessment task included a practical science exploration of floating and sinking. This task included elements of both differentiation by outcome and differentiation by task. This description will focus on the central section of the task, which addresses some statements of attainment at both levels 2 and 3. These statements of attainment are set out in Figure 3.

This task made considerable organisational demands on teachers, and acquired some notoriety in the national press as a result. Before starting to assess a group of four children, the teacher had to gather together the following resources:

- four objects for each child, varying in weight, texture, shape, size and buoyancy, for example, a stone, an apple, a twig and a shell
- weighing scales
- a container of water

Figure 3: Statements of Attainment Addressed by the 'Floating and Sinking' Task

<p>Level 2</p> <p>Sc 1/2f <i>Record findings in charts, drawings and other appropriate forms</i></p> <p>Sc 1/2c <i>Use non-standard and standard measures</i></p> <p>Sc 1/2d <i>List and collate observations</i></p> <p>Level 2</p> <p>Sc 1/2a <i>Ask questions and suggest ideas of the 'how, 'why, and what will happen if variety</i></p> <p>OR</p> <p>Level 3</p> <p>Sc 1/3a <i>Formulate hypotheses</i></p> <p>Level 2</p> <p>Sc 1/2b <i>Identify simple differences</i></p> <p>OR</p> <p>Level 3</p> <p>Sc 1/3d <i>Select and use simple instruments to enhance observations</i></p> <p>Level 2</p> <p>Sc 1/2e <i>Interpret findings by associating one factor with another</i></p> <p>OR</p> <p>Level 3</p> <p>Sc Sc 1/3h <i>Interpret observations in terms of a generalised statement</i></p>

- hand lenses
- writing equipment
- a recording sheet supplied as part of the standard assessment task.

Watched by the teacher, each child in the group then had to draw each of his or her objects, weigh it, make a prediction as to whether it would float or sink and say why, and then test out the prediction and record all the results.

Some of the assessments were relatively straightforward to make: for example, the ability to weigh and to complete the recording sheet with reasonable accuracy. Others, though, required some scientific

understanding on the part of the observing teacher, for example in distinguishing between predicting and hypothesising.

Children were generally enthusiastic about this task. One little girl observed on a case study visit said 'I love this work. I want to do it again and again and again.' One example of a completed recording sheet, reproduced as Figure 4, [to be included] perhaps gives a flavour of the interest and involvement in the task, with its careful observations and precise, lively descriptions of each object's behaviour in water.

Teachers, on the other hand, were less enthusiastic, though their reactions were mixed rather than entirely negative. Again, as with the dice game in mathematics, this task required constant, intensive concentration on a small group of children. But in addition to this, the task itself was time-consuming, taking at least an hour even at a brisk pace. It is important to realise, too, the organisational implications for teachers with large classes, who had to repeat the task seven, eight or nine times in order to assess all the children in small groups. Many teachers welcomed the practical nature of the activities and appreciated their pupils' involvement and enjoyment, but nevertheless condemned the task because of the excessive workload it entailed.

Assessing Reading

English attainment target 2 sets out the broad definition of reading in the National Curriculum and the attainments required at levels 1, 2 and 3. At level 1, only emergent reading is assessed; at level 2 a number of skills and strategies and the capacity to read a range of straightforward texts; at level 3, silent reading and a more sophisticated understanding of content. The statements of attainment at these levels are set out in Figure 5.

The reading task at each of the three levels consisted of individual reading and a discussion with the teacher. Each level featured an appropriate choice of text, and an assessment approach carefully matched to the statements of attainment.

At all three levels, the reading assessment was based on good quality children's books of the kind normally found in class book corners, not on graded reading schemes or specially constructed test passages. The National Curriculum programmes of study place great emphasis on the importance of providing a 'range of rich and stimulating texts' for children to read. Offering a selection of real literature was an important feature of the tasks, and contributed to validity both in reflecting the curriculum and in motivating the children to show their best performance.

To provide consistency of assessment, however, a new method of arriving at comparability of difficulty level of these books had to be developed. A readability measure was just one of the five elements considered. The others were, for level 2, the support offered by illustrations; line length; repetition and rhythm; and familiarisation. At level 3 the readability score was supplemented by measures for line length, support from illustrations, the

total length of the book (as sustained silent reading was required), and the scope for understanding 'beyond the literal'.

The level 1 assessment was a discussion with the child about a familiar book, in the course of which a few words and letters were identified. At level 3, the child read a whole story silently and then talked to the teacher about the plot, characters and setting, to show whether he or she could infer meanings, reasons and motivations and understand the structure of the story.

At level 2, however, the statements of attainment gave rise to a particularly distinctive assessment approach, based on a modified miscue analysis. As the child read a passage of about 100 words, the teacher was asked to make a running record of the child's attempt at each word, noting not just those read correctly, but also substitutions, omissions, phonic attempts and words that had to be supplied by the teacher. An example of a completed running record is given as Figure 6. [To be included] This record provided evidence for the teacher to make assessments of accuracy and independence, and of the ability to make use of a variety of cues in reading.

The approach to reading was well received by both teachers and children, and it appears that many teachers developed their expertise in diagnostic assessment as a result. There were some reservations, however. There was again the problem of manageability, as teachers of large classes had to fit in thirty or more lengthy individual interviews. The painstaking analysis of the difficulty of the texts was not explained in detail to teachers, and many complained that the books varied in difficulty, an opinion that appeared to be based solely on the vocabulary used.

Overall, however, this assessment was particularly welcomed for its obvious fulfilment of the formative and professional development purposes of National Curriculum assessment.

Conclusion

In his opening paper, Chris Whetton identified the main themes of this symposium as the validity, reliability and manageability of the National Curriculum assessment system, and the interplay between these three themes. This more detailed description of some of the standard assessment tasks has illustrated how these three aspects were embodied in the tasks taken by half a million children in 1991.

In the number task, the worksheets provided a valid, reliable and manageable assessment, but only of a very limited area of subject-matter. The dice game assessment of number facts raised several questions about reliability. This highlights an important feature of the National Curriculum. It does not restrict itself to easily defined, easily assessed pieces of knowledge, but aspires to set out a complete, broad definition of each subject. This was particularly evident in the practical science task, where high curricular validity was accompanied by serious problems of manageability and some doubts about reliability.

Figure 5: Statements of Attainment in Individual Reading Assessments

En 2/1a	<i>Recognise that print is used to carry meaning, in books and in other forms in the everyday world</i>
En 2/lb	<i>Begin to recognise individual words or letters infamiliar contexts</i>
En 2/1c	<i>Show signs of a developing interest in reading</i>
En 2/1d	<i>Talk in simple terms about the content of stories, or information in non-fiction books</i>
En 2/2a	<i>Read accurately and understand straightforward signs, labels and notices</i>
En 2/2c	<i>Use picture and context cues, words recognised on sight and phonic cues in reading</i>
En 2/2f	<i>Read a range of materials with some independence, fluency, accuracy and understanding</i>
En 2/3a	<i>Read aloud from familiar stories and poems fluently and with appopriate expression</i>
En 2/3b	<i>Read silently and with sustained concentration</i>
En 2/3d	<i>Demonstrate, in talking about stories and poems, that they are beginning to use inference, deduction and previous reading experience to find and appreciate meanings beyond the literal</i>

In reading, the success of the task was again its validity, but it was extremely time-consuming and gave rise to some complaints about manageability., Reading is, of course, an area where conventional standardised tests abound, and, not surprisingly, the reliability of the reading assessment has been questioned by proponents' of such tests. These comparisons are not entirely fair, in the light of the innovative criterion-referenced system and broad definition of reading addressed by the standard assessment task.

Overall, as Chris Whetton suggested, the success or failure of the system in 1991 can only be judged in the light of one's own perspective. The mixed reactions to the tasks described in this paper, and the range of perspectives giving rise to those reactions, typify the responses amongst education professionals in England and Wales in 1991.

Supporting the Teachers

Steve Hopkins Bishop

Grosseteste College

Introduction

It was recognised at an early stage that the successful implementation of National Curriculum Assessment would be dependent on the quality of the support for the teachers who were required to operate the system. The purpose of this paper is to provide an account of: the needs which the various form of support provided were designed to address; the nature of the forms of support themselves; and the perceived effectiveness of the support as judged by the teachers.

What Needs was the Support Designed to Address?

Knowledge of the system

In the first year of operation considerable emphasis was given to helping teachers come to terms with the system of attainment targets and the associated statements of attainment and levels. Coming to terms with the system involved, at the surface level, knowledge of the new vocabulary for its components and at a more fundamental level with, the nature and implications of a criterion-referenced rather than the more familiar norm-referenced model.

Making Use of Assessments for Formative Purposes

One of the purposes of the new system is that the information provided should be used by the teachers to plan the next steps in learning for individual children. As teachers become familiar with the workings of the new system emphasis is increasingly being placed on helping them to use assessments as a tool to enhance progression in learning and to draw attention to the need for differentiated teaching.

Helping Teachers to Develop Skills in Assessment, Recording and Monitoring

The assessment of pupils' progress is not new to teachers. The demands of a criterion-referenced system requiring the assessment of specific competencies during 'normal' classroom activities highlighted, however, the need for development work related to teachers' skills in assessing and recording pupil achievement. This was seen as crucial if teachers were to use assessment information formatively.

Helping Teachers to agree Standards

The statements of attainment which express the knowledge and skills which pupils are required to attain in order to be deemed to be at a particular level are, in many cases, open to interpretation. Because of this it has been necessary to provide teachers with a variety of forms of support which seek to enable them to come to an

agreement about the standards of work which particular statements of attainment represent. *Agreement trialling* is the term used to refer to a range of activities in which teachers engage in order to reach consistent interpretations of statements of attainment and the standards expected.

Helping Teachers to Use Standard Tasks Effectively

It has been necessary to help teachers to use standard tasks effectively in order to ensure, as far as possible, the reliability of the reported outcomes. At the surface level it has been necessary to support teachers in finding their way through the standard task materials, and in understanding the procedures they are required to follow. At a more fundamental level, support has been required to help teachers to organise and manage the activities within the classroom setting. For many teachers the methods of working i.e., the classroom practice which the tasks demanded, whilst a valid reflection of what is widely recognised to be 'good' practice, was not in fact practice which was current within their classrooms. This together with the demands that would be made by the need to administer the tasks within a relatively short period of time indicated that support related to classroom management was going to be crucial.

What Forms of Support were Provided?

In order to meet the various needs identified above the following forms of support were provided: support in the form of printed materials; support in the form of training programmes; support in the form of telephone help-lines and self-help groups; and support in the form of visiting moderators.

Support from Printed Materials

A range of types of printed materials were provided for teachers dealing with aspects such as; the nature and purposes of the new system; using the results of assessments for formative purposes; the development of teacher skills in observing and recording evidence of attainment. Some of the materials were produced centrally by the government agency responsible for implementing the new system, other materials was produced by Local Education Authorities (the bodies responsible for administering the education system at the local level) and by other agencies. To support agreement trialling, a range of materials were produced centrally which provided teachers with examples of children's work (written work and descriptions of practical activities) together with commentaries which included the reasons why these examples were deemed to meet the requirements expressed by particular statements of attainment. Those areas of the curriculum for which the printed word was not an appropriate vehicle for communicating the standards expected e.g. reading, were supported using video extracts.

Support from Training Programmes

Central government provided funding for Local Education Authorities to set up appropriate Programmes for teachers to acquire the knowledge, skills and confidence needed to successfully implement the new

assessment arrangements. A cascade model was used with training being provided at the national level for LEA trainers who in turn delivered the training at the local level. In broad terms the training focussed on each of the needs previously identified with particular emphasis being placed on agreement trialling and the effective use of the standard tasks in the classroom.

Support from Help-lines and Self-help groups

In addition to the provision of training programmes many LEAs provided telephone help-lines for teachers and schools. These were used to solicit advice and guidance on particular aspects related to the administration of the standard tasks and to seek clarification on evidence of attainment statements. An equivalent telephone help-line service was also provided centrally for use by LEA trainers.

An additional form of support was provided, in many cases, by the teachers themselves. Across the country various groups of teachers met outside school time in a variety of self-initiated selfhelp groups.

Support from Visiting Moderators

Each of the LEAs were required to demonstrate an active quality assurance strategy in the interests of the consistency of the reported assessment results from schools across the country. A central feature of this strategy was the role played by LEA moderators. Each LEA appointed a number of moderators who visited schools to ensure that the standard activities were being administered in broadly equivalent ways.

The Effectiveness of the Support

Knowledge of the System

Most teachers would acknowledge that they now have a reasonably sound grasp of the requirements of the new system. They are working confidently and comfortably with a curriculum framed by attainment targets and they recognise the benefits which have ensued with regard to a more focussed consideration of children's achievements. Professional-development has been fostered especially with regard to increased skills in classroom management as necessitated by the need to observe children's reactions to activities carefully and in relation to the systematic recording of children's achievements.

As yet, the full benefits of a system designed to promote the use of assessment information formatively remain largely unrealised. Whilst great strides have been made in terms of assessment itself the use of assessment information to benefit teaching and learning has, perhaps understandably, taken a back seat. Teachers are, in the main, directing their energies towards the statutory requirements of the system, i.e., to make and report assessments.

Agreement Trialling

The various processes of agreement trialling have proved to be very effective vehicles for professional dialogue between teachers about approaches to teaching and learning and about expectations in relation to standards of work. Where teachers meet together within and across schools to discuss examples of work from their pupils and to locate features of the work within the context of examples provided nationally, the benefits have been marked. Some teachers have begun to extend the agreement trialling notion to aspects of the curriculum which are not so easily considered using examples of work on paper. In these cases teachers are devising tasks and agreeing observable outcomes which reflect the statements and the standards they embody.

Materials and Management

The majority of teachers found that the support they have received has been effective in helping them come to terms with the requirements of the standard tasks. It would be fair to say, however, that with regard to the management of the tasks within the context of normal classroom practice the support provided did not really ease the difficulties experienced by many teachers. Whilst the tasks were reasonably valid reflections of good classroom practice the requirement that they should be administered in a relatively short period of time posed considerable problems. Even in cases where the current practice of a teacher was such that she/he could spend periods of time working, uninterrupted, with small groups of pupils (an arrangement required by the standard activities) the requirement that this arrangement continued over an extended period of time was problematic. Teachers felt that they were unable to give children the attention they required and that when they were working with a group on a standard task the quality of the curriculum for the others was compromised.

Whilst issues related to classroom organisation and management were anticipated and therefore addressed within training programmes and through other forms of support, the variation in the classroom contexts within which the teachers worked prevented any really significant impact being made. Teachers found themselves in somewhat of a dilemma. On the one hand they recognised the validity of the tasks themselves and wished for future statutory assessments to be based on similar tasks which reflect good curriculum practice rather than those based on more formal tests. On the other hand, however, the management difficulties which the tasks posed clearly indicated that it would not be possible for the approach to continue if the short time period in which they had to be administered remained.

By and large, the various strategies employed to support teachers were effective. The true effectiveness, however, was dependent upon the attitude of the teachers themselves to the change the new system was demanding. Those teachers who were helped to see the potential that would accrue for the quality of teaching and learning were more open-minded and receptive to the support offered. Those teachers who were sceptical or even hostile limited their own receptiveness to the help that was provided.

Despite the problems teachers encountered with regard to the management of the standard tasks in the classroom it should be noted that one year on, and with the benefits of a less pressurised atmosphere, many teachers are now acknowledging that the difficulties they experienced and the strategies they used to resolve them have had a lasting effect. Many teachers now organise their classrooms and manage teaching activities in such a way as to allow them to spend periods of uninterrupted time with groups and with individuals. Many have seen the advantages for themselves and for the children of encouraging and making practical arrangements to facilitate greater pupil autonomy.

Changes for 1992

Dorian Bradley

School Examinations and Assessment Council

The first ever national assessment of seven year old children in England and Wales took place in the early Summer 1991. Assessment tasks were distributed to 20,000 schools in February 1991 so that 40,000 teachers could be trained on their use, prior to the 6 week administration period of April and May. The reaction to the materials before during and after their use with 620,000 children set the scene for changes to the system for 1992

Once published and distributed to schools, there could be no formal mechanism for keeping the materials confidential as teachers needed to familiarise themselves with an innovative assessment package, to attend training sessions, to marshal resources and to plan quite carefully how to manage the six week assessment period. Samples of the material were therefore presented to the media. The resulting coverage tended to be cynical and based on an incomplete understanding of teacher mediated assessments. The package did not look in any way like the examinations and tests that the journalists remembered from their own school days. They looked for, but did not find many items that would have produced 'good copy', for example, sheets of sums, a list of words to be spoken correctly; essay titles; short objective tests based on recall of science facts. Many articles decried the absence of such approaches to testing. A week or two after the initial coverage in the national press, more balanced and informed articles from educational journalists started to appear but they did not get wide publicity nationally, as the media world had moved on to other issues by then.

The list of books on which the assessment of reading was to be based however was a source of many columns inches with little effort on the part of journalists. It therefore received prominent coverage in every national newspaper. This publication of 'approved national curriculum test books' led to very heavy purchasing of the titles by schools to ensure, that they had copies of every title, even though that was not necessary, and by parents, presumably to prepare their children for the test. Stocks soon became exhausted! This initial wave of publicity took place in January, a month before schools received the materials and before the start of the training programmes arranged for teachers during that Spring. It led to great uncertainty and anxiety amongst teachers and parents alike.

During February and March, SEAC received many letters criticising both the assessment model itself and the material prepared for teachers and pupils. These came from teachers, head teachers, governing bodies of schools, and parents. Much of the criticism was not founded on a reading of the material, but merely reflected what the letter writers had read, or very often what they thought they had read, in the newspapers or had heard on the television. As the training progressed during the Spring, some more thoughtful and balanced letters

arrived as teachers started to work with the materials in preparing for the administration period after the Easter holidays.

This break in the school year is the time when national conferences of the professional associations of teachers are held. Naturally enough, this first national testing of children was very prominent in their deliberations. The underlying theme of the debates was a philosophical opposition to an imposition of tests to be used with all children of a certain age group at a certain stage of their school life. While there was little opposition to assessment itself - 'teachers do it all the time - the clarion call was for a bank of standard assessment tasks that teachers could draw upon to use as and when it would be appropriate for individual children. This led to a second wave of unsettling publicity immediately prior to the assessment period.

Visits to schools by SEAC officers while the standard wsks were being used unearthed a number of areas of concern being expressed by teachers. These initial findings were confirmed by later evaluations. Five statements in particular came to dominate the aftermath of the 1991 assessment.

- 'the standard tasks didn't tell us anything we didn't already know'
- 'it was too much to do in too short a time and it was too repetitive in nature'
- 'what about the other children? I need adult support for them while I carry out the assessment of a small group'
- 'these tasks are not standardised, different teachers will do them in different ways'
- 'the levels are too broad to be of any use, especially level 2 in reading'

Changes for 1992

The planning for the 1992 assessment took place during June to September 1991. It had been recognised right from the first days of the development in October 1988 that it was unlikely that an innovation as great as national testing would be perfect in its first year, and that changes to the model would be inevitable. By July, sufficient evidence had emerged to show that considerable benefits had been brought about for both teachers and children by the use of the standard tasks. Consolidation, rather than whole-scale change, was therefore adopted as the basis of preparation for 1992.

Assessing Number

This attainment target seemed fertile ground for reducing the workload on teachers by replacing the dice game of 1991 by the provision of an arithmetic test. This would allow all children at any particular level to take the test at the same time.

Concerns about standardisation could be met by requiring all teachers to use the same numbers and by defining a time limit within which children had to make their response. Worksheets (Figure 1) [to be included] and the 'five second rule' (Figure 2 and 3) [to be included] is the approach adopted for 1992.

There have been some wry comments about using outlining drawings of fruit as location devices and the apparently arbitrary limit of five seconds. However, teachers have the freedom to prepare their own worksheets, to use flash cards, or to do the whole assessment orally with smaller groups of children. Once teachers have realised that this flexibility exists then any initial opposition to the approach evaporated.

The 'five second rule' has arisen from some limited trialling. It appears to be very generous if a child actually does know the number facts. If a child has to compute or count however it soon becomes apparent to the teacher that the child is not demonstrating success against the statement attainment.

Assessing Exploration of Science

The floating and sinking activity, along with a practical assignment to assess Using and Applying Mathematics, had easily been the most time consuming part of the 1991 assessment. Neither of them features in 1992, although teachers will still need to make their own Teacher Assessment of these parts of the curriculum. This highlights a fascinating dilemma for educators dealing with mathematics and science. Inclusion of these process based attainment targets in a testing system using standard material in a closed, rather than open, investigation is seen by some as a perversion of the nature of these parts of the subject. However, not including them could reduce the importance of process skills in the eyes of teachers. This gives rise to fears that teachers will develop teaching programmes concentrating on knowledge of scientific fact, for example, to the detriment of exploration and investigation.

Along with the statutory material, a number of optional standard tasks were published and distributed to schools before Christmas 1991. Each one focuses on an individual attainment target. Teachers have the option of using them if they wish. Some visits to schools early in the testing period shown that many teachers intend to use the new optional tasks written for Exploration of Science (or indeed the floating and sinking activity from 1991) with one or two groups of children, to fix a standard when they make decisions about the attainment of the rest of the children in the class.

The broad shape of the 1992 approach was set out in a major speech on education by the Prime Minister in July. In a section devoted to national testing he said

“tests are essential, And tests are here to stay. Of course, tests are not the be-all and end-all of education. We are not in the business of putting pressure on our children. But we must be able to measure their progress in an objective and regular manner.

Tests for seven year olds must deal precisely with the core skills, the "three R's". We need to know where things are going wrong. at an early age - or we will never be able to put them right.

It is early days - and I readily accept that we may not have got the process right yet. Where it is wrong, we will change it. Testing must not dominate the classroom. It must not swamp schools in paperwork. Nor should it be driven by too theoretical an approach. We need to shift the emphasis toward shorter standardised tests, which the whole class can take one at a time"

I shall deal with the implications of working these principles through into practice for the three activities assessing number, exploration of science and reading later in my talk. Some decisions about the regulations governing the assessment were being taken, primarily to reduce the burden on teachers, especially those with lap classes.

The final date for the completion of Teacher Assessment in 1992 was to be the mid point of the Summer term not the end of March as in 1991. As well as allowing more time to cover the curriculum before making decisions about attainment, teachers would also be able to use the standard tasks for particular attainment targets to help them decide both about the levels attained in them and in other related ones.

Teachers could start to use the standard tasks earlier, after the mid point of the Spring term rather than waiting until after the Easter break. This would provide a 13 week period compared with the six weeks available in 1991.

The number of attainment targets covered would be reduced from nine to seven.

Whenever possible the tasks would be designed for use with relatively large groups of children, all those in a class, or all those on a particular level.

Material for level 4 would be produced to help teachers measure the attainment of very bright 7 year olds.

Looking back over the last 12 months some teachers are now a little embarrassed by the outcry against floating and sinking. One teacher told me recently that she now realises that the activity allowed her to learn more about children than anything else she had done in thirty years of teaching!

Assessing Reading

The number of books on the list for level 2 has been reduced for two reasons.

Firstly, teachers had identified that the books on the list for 1991 represented a range of difficulty as measured against readability criteria such as the Spache index. Some books were highlighted as deviating markedly from the norm established by the majority of books. Commentators however did not take into account the other measure identified by the developers. Nevertheless it was decided to reduce the number of books on the list to increase the degree of comparability amongst the remainder.

Secondly, the government had commissioned the developers, through SEAC, to find a means of discriminating within level 1. Each passage used in the 1991 assessment had been marked on a 'running record' (Fig 4) [to be included]. A hundred or so of these grids were collected for each book and

subjected to computer analysis. For most of them it proved to be possible to identify 25 key words that the level 2 population into five roughly equal divisions ranging from a A grade for a perfect performance to an E grade for getting no more than 18 of the 25 words correct. The grade boundaries vary from book to book (Fig.5) [to be included].

Where these two matched for a particular book, it was placed on the list for 1992. This reduced the 27 books used in 1991 to 12.

The government also requested the development of a written test of reading. Such a test will be available on an optional basis for level 2 readers in 1992. It consists of a number of item types including a cloze procedure and a conventional comprehension test based on a short story. Where teachers choose to use this, it will place a child into one of five bands spanning level 2. The relationship between the two divisions of level 2 will be considered very closely when data becomes available in the Summer.

The only new books in 1992 are those for the assessment of children reaching level 4, the attainment level of an average 11 year old. Great care was taken to ensure the availability of these books and there has been no difficulty in supply this year.

Conclusion

The 1991 assessment was carried out in a period of extreme uncertainty in schools. It proved to be a hard going for the majority of teachers although children thoroughly enjoyed the activities and the close attention of teachers. The very detailed evaluations indicated where improvements in the system could be made for 1992 and this paper demonstrates that a number of significant changes have been made. Are they enough? Time alone will tell, but to date there has been very little criticism from teachers of the 1992 assessment arrangements or of the standard tasks.

Looking at the statements made by teachers after the 1991 assessment we can see that there is less to do and twice as much time available; the time consuming and repetitive tasks have been removed; increased provision of worksheets gives greater standardisation and allows more children to take part at any one time, although teachers have retained the flexibility to work orally with smaller groups; and there are two methods of determining a finer grading within level 2 reading. What about the first statement however, which shows that teachers felt that they had not learnt anything new about the children?

It took a little while to fully understand this, as in one third of cases, the Teacher Assessment levels recorded by teachers in March 1991 were different from those determined by the use of the standard task in the following six weeks. Our conclusion was that although the standard task in some cases might not have revealed anything new about a child, it served an extremely useful function in exemplifying that child's attainment in relation to the

National Curriculum. In other words, the tasks defined the standard needed to meet the criteria set out in the statements of attainment.

Independent evaluations are now also showing that teachers' knowledge of classroom management and assessment techniques also increased considerably as a result of the 1991 assessment. Teachers themselves are much less apprehensive than a year ago and very few critical letters have been received. It is with some confidence therefore that we look forward to this year's exercise.

Before concluding, there are a number of other developments that I should mention. This year, journalists will be able to find a spelling test! Children reaching level 3 or 4 in spelling during the course of their own continuous writing will be required to take a spelling test. Teachers will read a passage of about 150 words to the children who will each have a copy of the passage in front of them, but with 24 of the words missing. The test will consist of the children writing in the missing words when the teacher reads them. Twelve of the words have been chosen to match the level 3 statements of attainment and twelve the level 4. The test will give three grades for each level.

Other subjects are coming on stream. In addition to the core subjects of mathematics, science and English, teachers will also have to assess technology with history and geography joining the framework in 1993.

There is also evidence of interesting work being done on the data that the 1991 exercise produced. Although they are not considered to be wholly reliable as they arise from the first year of operation, many LEAs are beginning to identify trends at least and are talking about these as being factors in allocating resources. This interesting development will be monitored very closely. It might be worth an entire symposium in next year's conference!

The Development of Tests for 14-Year-Olds

Alan Greig

School Examinations and Assessment Council

In March 1989, three months after development work on standard assessment tasks for seven year olds had started, the government issued a specification for the development of standard assessment tasks for 14 year Olds. Work began in June of that year on tasks in mathematics, science, English, Welsh and technology.

The specification was, in many ways, similar to that for seven year old. Agencies were required to develop assessment tasks of the kind proposed by the Task Group on Assessment and Testing (TGAI) that is, packages of tasks administered through a range of modes of presentation (ways of giving the task to the pupil), operation (the way in which the pupil works on the task) and response (the way in which the pupil responds to the task). There were, however, two key features of the work which were to become important.

First, the tasks were to be designed to ascribe a pupils performance to any one of the 10 levels on the national curriculum scale as opposed to the three required for seven year olds. This had major implications for the design of the tests where, as has already been described, some of the attainment targets were structured so as to facilitate differentiation by outcome and some were more appropriate to differentiation by task.

Second, the tasks were to

- 'reliably and validly assess a number of attainment targets';
- 'motivate pupils and engage their interests';
- 'be easily administered, assessed and recorded by teachers'; and
- include 'evidence from written tests'.

It was not going to be a simple matter to satisfy all of these requirements at the same time.

The requirements of many of the attainment targets meant that a valid assessment could only be achieved by methods other than written testing. The need to motivate children pointed in the direction of classroom based assessment by task. The need for reliability and manageability pointed more in the direction of written tests.

By the spring of 1992 the work was ready to be piloted in a 2% sample (about 100) of schools randomly selected from England and Wales. The table below summarises the assessment structures used in the summer of 1991.

Subject	Type of Test						
	Oral	Aural	Practical	Short-written		Extended-written	
				Timed	Untimed	Timed	Untimed
English	✓	✓	✓			✓	✓
Welsh	✓	✓	✓	✓	✓	✓	
Mathematics			✓		✓		✓
Science			✓	✓	✓		✓
Technology			✓	✓	✓		✓

The exercise was fully evaluated and a report produced by SEACs Evaluation and Monitoring Unit (EMU). The full report is contained in 'National Curriculum Assessment at Key Stage Mute - A review of the 1991 pilots with implications for 1992'.

As far as manageability was concerned evidence suggested that while certain elements of the exercise were difficult to manage there was little evidence of large-scale management difficulties. Making assessments in the classroom proved to be difficult for many teachers, marking time was greater than normal and there were some difficult whole-school, management issues to solve. On the other hand, teacher preparation time was no more than would normally be needed, pupils were not stressed by the exercise and generally enjoyed their work during the pilot.

The validity of the tasks was evaluated in terms of their descriptive and construct validity. Descriptive validity was evaluated by the agencies in a number of different ways making it difficult to look comparatively across subjects. In general, however, panels of expert validators were asked to confirm that the tasks assessed what they were intended to assess. In one exercise the reverse happened and validators were asked to ascribe tasks to particular statements of attainment. Lessons were learned about the difficulties of measuring descriptive validity quantitatively. The exercise yielded variability of outcomes depending on the method of evaluation used by the agency, the statement of attainment being targeted and the expertise of the validator. A clear message for the future was that the tasks should be thoroughly pre-tested in order to identify and exclude inadequate items.

In construct validity terms the assessment materials were evaluated according to the extent to which they measured the key constructs of strands (hierarchical groups of statements of attainment), attainment targets and subjects. In English high levels of construct validity pointed up the very close relationship between the assessment tasks and teachers' normal work. In mathematics it appeared that the standard task offered a very

good structure within which to assess hierarchical strands (as found in 'process' attainment targets) but that there was much more uncertainty in the non-hierarchical areas (as in the 'content' attainment targets).

It was subsequently decided, as a result of the reorganisation of the mathematics and science attainment targets, that 1992 would be a voluntary national pilot exercise. At this time some 4246 schools have decided to take part. This figure represents about 76% of schools with 14 year-olds in England and Wales.

The tests will take place on 8 and 9 June and will comprise 3 one hour written tests in each subject. They have been designed to cover the majority of the attainment targets - three out of the four in science and four out of the five in mathematics. In each case the attainment target omitted is in the 'process' attainment - using and applying mathematics and scientific investigation. In both subjects the tests are provided in four bands of overlapping levels 1-4, 3-6, 5-8 and 7-10, thus allowing teachers to choose the band that suits the individual needs of their pupils.

It is likely that a pupils' subject score will be based on the average of the test results and the Teacher Assessment of the attainment target not covered by the tests. In future years the government may require the publication of these results.

Another feature of the 1992 assessment has been the introduction of a 'quality audit of the assessments. The five GCSE examining groups in England and Wales have been given contracts to check on teachers marking. In 1992 about 2250 schools will take part in this exercise which will have the prime aim of ensuring that the marking is consistent within and between schools. The 1992 quality audit is seen as important in ensuring that the test results are reliable and credible.

The shift from classroom based assessments to conventional tests reflects the government's concerns over manageability and reliability. It is also evident that the tests are more likely to fulfil summative, evaluative and informative purposes rather than to provide information that enables teachers to plan the next stages of a pupils work (formative purposes) or to assist in the professional development of teachers.

Consequences of the Changes

Teacher Assessment

SEAC has been concerned to emphasise the importance of teachers' own assessments as a means of ensuring that the formative and professional development purposes of assessment are met. Material has been produced to support teacher assessment of Ma1 and Sc1 (the process attainment targets), anthologies of assessed pupils work are designed to give some guidance on standards to be expected and publications of case studies of teacher assessment in the classroom aim to demonstrate examples of good practice.

High reliability of assessment tasks is essential if the results can be interpreted and used with confidence. It was always going to be difficult to achieve and demonstrate high levels of reliability. The agencies all carried out

re-marking exercises which generally gave rise to differences between marker and re-marker of one level. The EMU report concluded that given the developmental nature of the work 'there is no reason to believe that they (the tests) will not achieve levels comparable with other forms of examination'.

Two further features of the 1991 assessment deserve particular mention - the performance of children with special educational needs and the comparative performance of boys and girls.

The table below summarises and compares the mean performance of children with special educational needs (SEN) with those who have no such needs (Non SEN) in three attainment targets - En 3 (Writing), Ma 3 (Number) and Sc I (Exploration of science).

	En 3	Ma 3	Sc I
SEN	3.6	3.2	3.4
Non SEN	5.4	4.7	4.5

As expected the performance of children with special educational need is lower. What is Surprising is that the SEN group managed to achieve as much as about 70% of the success of the non SEN group- This appears to support the view that classroom based assessments work particularly well with these children. Reasons for this might include the fact that teachers were able to modify the tasks to suit particular children and were able to administer the tasks flexibly by allowing different modes of response over a flexible period of time.

Girls generally attained higher levels than boys - by about 0.5 of a level in English, 0.2 in mathematics and between 0.1 and 0.3 in science. With the exception of mathematics these findings are consistent with differences in gender performance in other examinations. The mathematics results were surprising since at GCSE (taken at age 16) boys do much better than girls. One possible reason for this difference might be the methods of classroom based assessment used in the mathematics tasks in 1991.

Changes for 1992

Before the pilot exercise could begin, however, the government required a shift in direction. The first national tests for 14-year-olds in 1992 were to be much more straightforward and sharply focused than the tests developed for 1991. The tests were to be simple, rigorous and objective and easy to administer. The model was to be one of time-limited examinations undertaken by all 14-year-olds simultaneously under controlled conditions with the scripts marked by the pupils own teachers.

Validity

Short written tests cannot assess the full breadth of some of the statements of attainment. For example, one of the mathematics statements covered in the tests is:

Calculate with fractions, decimals, percentages or ratio as appropriate. (Ma 216a).

This statement is to be found in the attainment target concerned with number and is placed at level 6. A short written test will not be able to assess each aspect of this statement. The 1992 test will focus on the calculation of percentages.

On the other hand quite complex statements can be assessed in a short test:

Demonstrate that they know and can use the formulae for finding the areas and circumferences of circles. (Ma 416d).

This statement is in the shape and space attainment target and is also at level 6. Test questions can quite easily be designed which require both the knowledge and use of the formulae for the circle.

Another feature of validity likely to cause difficulty in 1992 results from the banding model adopted. Pupils are likely to demonstrate uneven performance across the attainment targets simply because, for example, they are not as good at algebra as they are at data handling. In other instances it will be because the national curriculum is new and their teachers will not have covered the full programme of study. Such pupils may find that they are not able to show evidence of particular aspects of a subject within a test on which they perform generally adequately.

In pre-tests for the 1992 pilot exercises the following proportions of pupils were not able to be allocated a level:

Mathematics: AT2 - 0%, A13 - 1%, AT4 - 5%, AT5 - 7%.

Science: AT2 - 19%, AT3 - 5%, AT4 - 18%

The higher numbers of pupils 'dropping off the bottom' of a band in science is a reflection of the unfamiliarity of aspects of the science curriculum.

Manageability

The test developers have been required to keep marking times within reasonable limits. The evaluation of the 1992 exercise will indicate whether this has been achieved. However, as far as the individual pupil is concerned, there will be 6 hours of testing in 1992. This will increase to 12 1/2 hours in 1993 and 16 1/2 in 1994 as further subjects are increased. It is likely that this programme will require two complete weeks of testing in 1994 and subsequent years. This feature too will need to be carefully evaluated in order to assure all concerned that Pupils are not being tested to an unreasonable degree.

Summary

The model for national testing of 14-year-olds in England and Wales has been modified as ministers have clarified the requirements. The evaluation of classroom based assessment systems in 1991 indicated a qualified success for such systems particularly with children with special educational needs and, in mathematics, for girls. The 1992 system of testing has been designed to be more manageable and reliable and

yet may have consequences for validity. The overall manageability of the exercise, will need to be closely monitored to ensure that both teachers and pupils are able to cope with what is being asked of them.

Advice to U.S. Systems Contemplating Performance Assessment

Chris Whetton

National Foundation for Educational Research

After what we have heard it might be thought that the advice to anyone contemplating a national or state assessment system based on performance: assessment would be the same as that given to the young person contemplating marriage - *'Don't'*. However, to take such a negative view would be to ignore some of the very positive features that occurred as a result of the first run of National Assessment in England.

I began earlier paper by setting out the purposes of National Curriculum Assessment. To reiterate, these were to be

1. formative - information for next stages
2. summative - overall information on achievement
3. evaluative - information on classes and schools
4. informative - information for parents for professional development

There is no doubt that the final purpose, has been an outstanding success. Teachers of young children had little or no knowledge of assessment or testing. The national curriculum itself was new and contained elements that had not previously been widely taught; for example the process elements of using mathematics and exploring science but also many of the knowledge elements of science. Through having to carry out the assessments and having to use the standard assessment tasks, teachers, became much more familiar with the curriculum and its assessment. They now understand its requirements to a much greater extent than teachers of older primary children who have not yet had to carry out the compulsory assessments.

The assessments had assumed a model of good practice in which teachers were able to deal with a small group of children and at the same time continue to ensure that the remainder of the class were continuing to learn. In fact, this proved to be very difficult for teachers to do. However, the experience has caused them to look closely at their teaching methods and to reflect on whether they were in fact engaging with children often enough.

Interestingly, such experiences have also become part of a debate on primary teaching methods; such a debate would almost certainly have occurred in any case but the realisation that the form of assessment implied by the currently generally advocated teaching style had caused difficulties came as a profound shock to some.

The system also began to meet some other of its purposes; the need to record pupil progress, meant that teachers began to consider the match of their teaching to their children - the formative purpose. In some

schools, results were kept for the first time and parents received reports on the children's progress. In a sense, the accountability of teachers to their children's parents became formalised for the first time.

Some other experiences were less happy. The national results were published and interpreted in a way which was misleading and detrimental to teachers and some education authorities. At the same time, those who favoured standardised norm-referenced tests complained about the reliability of the results. It is certainly true that in the first year of a system based largely on professional judgement, in which teachers have received differing amounts and quality of training, will not have led to uniformity of standards. Such uniformity can only develop over a period of time, and the interpretation of results needs to reflect this.

This brings me back to the balance which I discussed between validity, reliability and manageability. It is evident that the different purposes imply different balances between these. Formative assessment implies highly valid classroom-based assessments, but that may not be summative and evaluative purposes generally require the balance to be toward reliability, but validity may be sacrificed. The informative and professional development purposes require both validity and reliability, but these may only be achievable over a long period of time. Also, the exact way in which validity and reliability are achieved is dependent on the particular attainment target, and hence as I said before the choices between them become important in determining the manageability of the formal assessments within the system. Hence, while the system as a whole may aspire to fulfil all its purposes, its various elements may meet only some of them. Ultimately, Teacher Assessment will largely fill the formative purpose. However, only through the experience of completing the standard tasks would professional development be sufficient for teachers to do this unaided. If possible, the roles of the different elements in meeting the purposes should be made clear.

Other purposes remain to be refined - the informative function of providing information to parents must also rely on a mixture of Teacher Assessment and the standard task, again emphasising the integrated nature of the system. As we have heard, the reaction of some parents and teachers was that the grading system, based as it was on only three of the ten levels, was not sufficiently informative. For reading, this has meant the provision of extra information within level 2 to provide greater differentiation. For all attainment targets, the assessment will now take place at the next level as well. Such criticism was based on a narrow view of the system as a whole which should be regarded as applying to children throughout their time in school. This progressive, integrated structure of levels is one of the glories of the system. However, because the assessment of seven-year-olds was the first part of the system to be introduced, this great advantage has been overlooked or misunderstood. It should not be forgotten that any entirely new system has to be sold not just to teachers but also to parents and the public.

What we have been describing has not simply been about assessment, The integration of curriculum, assessment and teaching methods should have been apparent. In fact, the assessment system has been one of the major forces in driving along change in schools. There have been complaints that the pace of change has

been too great. It has also been suggested that there should have been much greater experiment and trialling before the implementation of the assessment system. In answer to this, I would suggest that for many years before 1988, education had its experiments, its trials and its local implementation but did not fundamentally change. Since 1988, the pressure for change has been considerable. In implementation, some that all of us mistakes have been made and corrections are occurring. This is surely the way work most of the time. We do not expect to instantly have a perfect book written, an impeccable garden created or fully efficient business running. We do our best with a first attempt then, learning as we go, improve matters redirecting our efforts accordingly. It is this type of model of change which has been disturbing for some. To quote Michael Fallon 'Pressure without support leads to resistance and alienation; support without pressure leads to drift and waste of resources'. We have moved from the latter situation and for some infant teachers have reached the former. However, for the majority, the pressure has challenged their professionalism so that real change has taken place. Also, because it has involved all schools and all children, real change has taken place in a short time throughout the country. Such a situation could not have occurred without some changes of direction during implementation.

I would sum up the messages I would give to a US audience contemplating state or national assessment incorporating performance assessment as follows:

- specify the system as a whole, making clear what its aims are and how its elements meet these aims.
- be prepared to make changes to this specification, and prepare others for the probability of change.
- carefully consider the balance implied by the aims of validity (curriculum authenticity), reliability and manageability.
- do not assume that one model of classroom practice holds for all schools and ensure that the assessments are appropriate for a diversity of practices.
- educate the various constituencies (teachers, parents, public, politicians) in the aims of the system and ensure that their priorities are being met.
- attempt full-scale implementation rapidly.
- provide a balance of pressure and support

To conclude, I would emphasise that it 'is possible to introduce major changes to an education system in a short period of time, that performance assessment can have a leading role in implementing curriculum change and that this can lead to a raising of standards. However, this is not an easy option, implementation will be difficult and your aims will be misunderstood.

Symposium II

Alternative Assessments in Practice: Perspectives on Issues and Problems

The intensity of the performance assessment movement led Joe Hanson to organize this symposium on the problems of the practitioner in dealing with its facets. Paul LeMahieu and Joanne Eresh presented innovations they incorporated into Pittsburgh's portfolio process. Maryellen Donahue suggested a metaphor of a performance assessment train out of control because there was no engineer. Discussions were provided by Judy Arter and Peter Wolmut

The Portfolio Reporting Project

Paul LeMahieu

JoAnne T. Eresh

Pittsburgh Public Schools

Even a cursory review of educational headlines in America over the last year reveals the level of dissatisfaction with the current means of assessing student progress, as well as a debate on the merit of alternatives. For example, 'Two Groups Laying Plans to Develop National Exams.' (Education Week, September 26, 1990); 'The Nations Report Card Goes Home'. (Phi Delta Kappan, October 1990): "Multiple Choice and It's Critics." (The College Board Review, Fall 1990): and "Advanced Competency." (The New York Times November 4, 1990) are only a few of the articles on the topic over the past year. What emerges is a persuasive case to reassess assessment, or to move from a "testing culture" that narrows teaching and distorts learning, to an "assessment culture" in which evaluation becomes an integral part of learning.

Such an examination is underway in the Pittsburgh Public Schools. Its PROPEL project (a loose acronym that incorporates Perception, Reflection, and Production) has been operating in the District's Writing, Visual Arts, and Music Classroom since 1987 and recently extended into science and mathematics. Teachers and students have been engaged in experimenting with portfolios as a way to integrate assessment into the learning process. The Pittsburgh portfolio is a collection of a student's work over time and across a range of instructional activities. The portfolios are not just collections of finished pieces (much less "best products"), but sketches, works in progress, reflections, self evaluations, successful and unsuccessful pieces as well as journal entries. They encourage students to be reflective about their efforts. The portfolio may subsume any or all of a student's best work, a history of particular projects, the processes a student undertakes, or the individual working style the student has acquired. For example, the student portfolios may include thumbnail sketches, further developed sketches, cartoons, notes from friends about their work, and ongoing written dialogues between teachers and students. Through a process of increasingly focused and personalized approaches, the materials in the portfolio gradually take shape.

As in many places, the most well developed portfolio work is occurring in the writing classrooms. A great deal of effort has been expended by teachers, supervisors, and administrators from the School District and researchers from Harvard and ETS to introduce, refine, expand and promote the use of portfolios as measures of student growth and development as writers. With this portfolio work now well established, the district is now developing and piloting the procedures for widespread and public accounting based on the reporting of information gleaned from portfolio assessment activities. This work will inform the current discussion of assessment both locally and nationally and will include input from key stakeholders both within and outside the circle of professional educators. With the development of valid alternative measures of student

accomplishment, educational reform moves one step closer to the challenging, intellectually honest, yet fair accountability that all parties recognize it needs.

If efforts at school reform are to take root and effect lasting change, traditional forms of testing will have to yield to models of assessment that can be genuinely subsumed as an integral part of the learning process. Whereas traditional testing is strongest at measuring the mastery of facts and the recall of information, authentic assessment is meant to provide opportunities for students and teachers to learn about the standards of good work with respect to more valued outcomes. The student may even be incorporated as an active agent in the evaluative process, not merely as an object to be evaluated. Moreover, traditional indirect methods of student assessment deliver influential messages, not only about what is valued as educational outcomes but the educational experiences and means of teaching that are most appropriate. Part of the motivation underlying direct and authentic measures of student accomplishment is the establishment of a resonance between the forms of assessment and the curricular and pedagogical approaches that we desire to support in the classroom.

The portfolio work that has evolved in Pittsburgh's writing classrooms recognizes this dynamic, dialogical nature of assessment. Portfolios are assembled and developed from student's writing folders that are used in all English and language arts classrooms in the district. From the middle of the school year on, students in grades 6 through 12 are asked to select from their writing folders a piece of writing they feel is important to them. Included with this piece are all the rough drafts and revisions, any comment sheets produced through discussion in a peer revision group or from their teacher, and the writer's explanation of why this particular piece has been chosen as important. These materials become the first entries into the portfolio. As the semester progresses, students are asked several times to return to their folders to select additional pieces for the portfolio and to explain why each piece exemplifies whatever judgment students have been asked to make: for example, why this is an "important" piece, or a "satisfying" or "unsatisfying" piece, or a "free pick" piece (one the student would like to include in the portfolio to "round out" the portrait of herself as a writer).

Thus, the portfolios created by the end of the school year are designed to give a complete portrait of the student as a writer. Students have been asked to become metacognitive regarding their own learning; they have been asked to use their entire collection of writing as a text from which to learn about themselves as writers and learners. Within this framework:

- assessment focuses student's attention on major ideas and significant processes rather than on final products alone.
- assessment is an ongoing process wherein students are able to practice and revise in order to accomplish good work:

- assessment is a collaborative. as well as an individual enterprise. i.e., students learn from peer reviews. they gain from reviewing samples of other students' work, they benefit from consulting with teachers and mentors,
- assessment is not just a chance to discover "how someone else thought you did," but an opportunity to learn about setting worthwhile and challenging goals, formulating and using appraisals, as well as an apprenticeship in self evaluation,
- assessment offers the opportunity to explore valuable learning outcomes that go well beyond that which falls within the purview of typical testing programs;
- assessment that also represents a meaningful and interesting challenge in itself and that provokes the strongest performance possible from students.

Traditional standardized testing has become the yardstick by which individual students, public school districts. and the education system as a whole are judged. The statistical standards known as the national norms allow parents. elected officials, professional educators. and students to know how a given school district performs in relation to other districts locally, nationally. and even internationally. However, such traditional methods of conducting assessment and making a public accounting lack the necessary resonance to permit them to support the kinds of classroom experiences that reform minded educators seek to promote. Though we are well aware of the tests that measure performance against the norms, there is no valid, credible, or reliable alternative measure by which the general public and other concerned parties can judge how well or how poorly a school district is performing.

On the other hand. few (some would say none) of the emerging programs of alternative assessments have undertaken the necessary exploration of their validity, reliability, and credibility that would permit them to support a good faith accounting to the public. While systems of portfolio assessment (Pittsburgh's included) do produce evidence in a form that is useful to support clinical judgements (e.g., determinations about students' strengths, weaknesses, and developmental needs). they do not routinely produce information that is easily summarized and aggregated in various ways so as to support a public accounting. This year Pittsburgh is exploring and developing such procedures. Of particular concern throughout is the determination of how such procedures might be employed without fundamentally transforming. even damaging, the essence of the portfolio process. In addition, the portfolio reporting project explores the ways in which a genuine "external or independent perspective" can be introduced into the assessment procedures,. so as to ensure the credibility of the public accounting. The objective, therefore. of the Portfolio Reporting Project is to demonstrate the feasibility and advisability of the use of portfolio assessment to make public accounting of the performance of the school system, of teacher meetings/workshops that include the derivation and evaluation of a representative

The project accomplishes its goals through three sets of activities. These include (1) a series sample of student portfolios from throughout the district; (2) the conduct of an external audit of the evaluation procedures and findings by independent reviewers: and (3) a reporting event at which the results of the accounting are made public. Each of these project activities is described briefly:

1. Teacher meetings/workshops A series of monthly meetings involving a number of teachers who have been involved in the development of the portfolio reporting procedures. These teachers span the grade range that is involved in the Portfolio Reporting Project (grades 6-12) so as to permit the portfolio assessment project to be carefully integrated with other direct assessments of student writing. The purpose of these meetings is to establish the guidelines and rules of selection for defining student portfolios, determine any modification necessary to existing scoring rubrics to permit the external accounting that is desired: monitor the assembly throughout the district of portfolios for all students, oversee the selection of a fair and representative sample of student portfolios. and at the conclusion of the year, rate that sample of portfolios to provide the necessary evaluation of the district.
2. Audit of portfolio evaluations by team of external reviewers One of the greatest challenges in the public reporting of alternative assessments generally, is the introduction of a necessary "external or independent perspective." This is necessary to ensure an honest and credible evaluation. In the case of traditional assessments. this independent perspective is introduced in two ways. First, the assessments are designed and developed independently of the agency being evaluated. Second, the explicit reference to national norms place the performance of the evaluated system in a (national) context that extends beyond its immediate circumstance. The challenge is to bring to bear the necessary independent perspective in a form of assessment that is locally defined and developed: locally administered: locally evaluated; and locally reported. One way of doing this is to constitute a team of external auditors. Typically this term might be comprised of members of the community. local businesses. and educational professionals from throughout the region. While the first demonstration project of this sort involves a fairly remarkable group of auditors (see list at end). the longer view is that this sort of professional review would be carried out routinely across districts. This external audit panel is trained with respect to the portfolio process: its purposes. goals, procedures and methods of evaluation and accounting. This training is conducted by the teachers and administrators from the school district implementing the project. The panel then rescores a representative. Sample of previously rated portfolios to verify the adequacy of the procedures and the fidelity of their application. The careful identification and selection of the membership of the external audit team will do much to

ensure its credibility as well as the credibility of the accounting that they will be called on the endorse.

3. Reporting event At the conclusion of the portfolio reporting project, a public reporting event is held. At it, an accounting to the public regarding the performance of the school system (as well as individual schools) will be made. Essential to the accounting is the presence of representatives of the external audit team who will certify that the statements regarding the performance of students in the Pittsburgh Public Schools accurately reflect assessments made using procedures applied and verified by them.

Researchers from Harvard Project Zero (HPZ) and Educational Testing Service (ETS) with whom the School District has worked closely since the inception of PROPEL, serve as consultants to the project. They advise the district on procedures for synthesizing and summarizing the information derived from the portfolios. Additionally, they consult on how best to approach technical issues concerning the reliability and validity of the data: how to aggregate data for public reporting; and how to amend the portfolio procedures to provide without damaging the educational processes that they are designed to support.

The Pittsburgh Public Schools have entered into a contract in a sense, to take a risk with an extended use of an alternative form of assessment. More significantly, it has undertaken this risk in a public forum. By attempting to develop a system of accounting collaboratively with teachers, and administrators as well as the public, the district recognizes that both real reform and real assessment necessarily involves all these parties. By considering its findings in conjunction with an outside audit team, the district makes an effort to avoid what one leader in the district has referred to as the "emperor's new clothes of testing." The development team believes it has a way to make assessment meaningful for students themselves, and, by means of this audit team, it asks the public whether these "clothes" are visible and meaningful to them as well. Such is the only gesture that will prove portfolio assessment's viability as a realistic and educationally sound vehicle for public accounting, as will give it justified standing as a useful component of our larger reform effort.

Performance Assessment: Implementation Issues

Maryellen Donahue
Boston Public Schools

Introduction

All Aboard! It is obvious that the performance assessment train has chugged out of the station and is steaming toward a far-off destination, unknown, yet powerful in its magnetism. What began as a leisurely foray has noticeably picked up speed and intensity to the extent that the surroundings are beginning to blur. That which originally seemed so dear is now muted.

The dining and day cars are filled with publishers and academicians, while the caboose spills over with school system testing and curriculum types. It is noteworthy that few teachers, students, and parents are on board. Could it be that most have been left at the station?

With regard to the engineer it is unclear who, or even if, there is one, or if she is equipped to handle the inevitable twists, turns, hills, and valleys as part of the journey. The rumor circulating in the back of the train and among the crowd at the station, however, is that there is no one at the wheel and that the riders had better be prepared for a perilous journey. Before damage is done, we need to examine what is going on and make adjustments. Then, we need to proceed cautiously.

The purpose of this paper is to describe some of the yellow and even red flags along the track and suggest ways in which implementation concerns can be addressed while still moving forward on this all important journey. The trip might require a number of changes, slowing down so that some people may catch up, moving in another direction, or even reversing temporarily. Very likely the journey will require new track on a more solid base to make the journey smoother and more stable, so that in the end we will be able to realize that the trip itself as well as the destination were well worth the effort. One wants to make sure that if the train goes into the tunnel, that it comes out the other end.

Broad Contextual Considerations

There are a number of over-arching contextual frameworks that need to be borne in mind in any discussion of assessment to minimize misuse or misinterpretation of assessment results.

One of the first considerations is determining the purpose for using a particular assessment. Given differing foci, there are varying sets of implementation issues. A test that is used at the local classroom level for diagnostic and prescriptive purposes can meet more relaxed psychometric and administrative requirements than either an evaluative assessment at the individual student level or at the level of a mandated large-scale assessment for accountability. Since most of the research and information has been in the area of individual

student diagnosis, as we move to the broader arena of accountability, there are likely to be additional issues that will arise; furthermore, if performance assessment is used for high-stakes accountability purposes, many of the same problems that were apparent with multiple choice tests will surface (Mehrans, 1992). To guard against this, we must be sure that assessments used for accountability face tougher requirements than ones used at the individual student level.

Other factors that need to be taken into account are the audiences for which the assessments are intended. Since the array of possible audiences is a large one that includes parents, students, classroom teachers, specialists, administrators and the community at large, and each of these groups has potentially different information needs, the specific language to describe the assessment tasks and results will vary. Descriptions of performance on various assessment tasks might not be relevant or understandable to all audiences. Furthermore, a lack of specificity in language will lead to confounding of the results.

The larger political context must be heeded as well. An instability or change in overall governance structure or a series of superintendents militates against clearly thought out and far-reaching assessment modifications. Politically, it is at the systemwide level that the accountability issues are most volatile and sensitive. The results of performance assessments not only must be clearly understood, but also must measure elements that the larger public deems important.

Another consideration is the inevitable power struggle between various interest groups, not only about what is measured and how, but also about varying information needs at the local, state and federal levels. Evaluation designs and requirements for most federal projects have not kept pace with the newer assessments, leaving a school district with limited options. Typically, in most large systems, individual schools have limited influence over what is to be taught and how it is to be tested systemwide. The nature of performance assessment requires that a delicate and negotiated balance will have to be attained so that the local school level control of testing and curriculum is not at odds with systemwide concerns. It is evident that existing roles and responsibilities of both teachers and administrators will be redefined considerably to accommodate the new dimensions of assessment

Particular Difficulties

There are a number of difficulties that, unless addressed, prevent the smooth and immediate implementation of performance assessment, particularly in large urban school systems. These problems are not insurmountable, but are generally understated in much of the literature on performance assessment. Unless they are dealt with and addressed somehow, they will severely distort the process and impair successful implementation.

One of the most prominent impediments that needs to be addressed is the issue of time. All aspects of the methodology are tremendously time-consuming: the instrument development, the administration of the assessment, and the scoring. In addition, teachers need common planning time to discuss standards and

measures of performance. As a reflective activity, both in development and in implementation, performance assessment requires time for adequate sharing and discussing of results or it becomes merely another testing activity of little relevance. Because of the overall change in the underlying theoretical structure, a lengthy time frame for development is required, one that extends over years.

It is important to note that such innovations in testing curriculum, and instruction cannot be thought of as distinct from teacher professional development. Any initiative to standardize curriculum or impose a particular assessment methodology should be evaluated in light of its probable effect on the teaching profession. (Ascher, 1990) The implementation of performance assessment is dependent on intensive and extensive teacher training and staff development. Thus, a massive restructuring of existing teacher education programs is required at both the in-service and pre-service levels. The challenges are heightened in systems with an aging staff with established teaching styles. Since the teaching approaches that long-term teachers have adopted are based on their own extensive experience and have become second nature, they are difficult to modify quickly.

The time demands and the requirement for extensive staff development are the key elements that combine to make performance assessment an extremely expensive model, which may be particularly difficult to justify in an era of diminished resources. While matrix sampling is sometimes proposed as a cost-saving measure, it runs counter to teachers' natural inclination to have as much information about each student as possible and thus it will meet with resistance at the classroom level. Since matrix sampling limits the amount of information teachers receive and subsequently lessens ownership of the assessment, some of the enthusiasm for the model is lost. Additional roadblocks occur because there are few articulated models and limited materials available for practitioners to review and react to; therefore, marketing the model to potential funders is problematic. Generally, representatives from external and internal funding sources such as the business community and city coffers require a concrete, specific, and practical proposal. Given the current developmental nature of this methodology fulfilling this requirement is difficult.

Change in Paradigm of Education

The focus on improved assessment mirrors other movements that are underway in education and are similarly based on the evolving disciplines of developmental and cognitive psychology that posit expanded theories of intelligences.

School restructuring efforts such as school-based management and shared-decision making are all predicated on new roles that emphasize increased empowerment and altered school structures. Within the classrooms, teachers are now being asked to take on the roles of mentors and coaches rather than classroom authority figures. Similarly, students are being asked to take on the role critics, critiquing their own work and, thus, move away from a passive role in their education. The arena is being broadened considerably to include teachers and principals not only as designers and developers of assessments, but also as scorers of assessments.

products. No longer is assessment the domain of only testing specialists, but it has been broadened to include collaborative and collegial investigation of alternatives.

Other trends such as the integration of curricular approaches both across subjects and across students with widely varying programmatic needs, cooperative learning, and the whole language approach all parallel the focus on integrated assessment and holistic approaches in curriculum and evaluation. As the emphasis is increasingly on the contextualizing of curriculum, so, too, is the emphasis in assessment on the meaningfulness of tasks for students and teachers.

The intent of the current testing movement is to convey something very different from the standardized instruments that were so heavily relied upon in previous educational reform initiatives, such as minimal competency in the 1970's and the standardized test-based accountability of the 1980's. This type of assessment requires a profound change in the structure and theoretical underpinnings of assessment. It requires new language and definitions in discussing the emerging models. Reliability and validity will need to have definitions, both academic and operational, that are very different from the past. Linn, Baker and Dunbar (1991) note that the traditional criteria of efficiency, reliability, and comparability need to be expanded at the same time that the forms of assessment are being expanded. It cannot be assumed that the burgeoning forms of assessment are any less prone to corruption than the traditional instruments.

Suggestions to Keep the Train on the Tracks

It is essential, then, to proceed deliberately. All activities at the classroom level that are aimed at improving assessments, their relevance, and their authenticity should be encouraged and supported. Great emphasis should be placed on performance assessment at the individual student level. At the classroom level, teachers intuitively respond to this type of assessment as being appropriate and valid. When it comes to high stakes testing such as accountability and competence certification, much additional technical work needs to be done in the areas of reliability and validity before school systems plunge full speed into a performance-based system to allow varied tasks to be piloted, revised, and improved over time without the complicating pressures from the political arena.

Thus with respect to the development of performance-based technology, such development must take place in the realm of low stakes. Large-scale collaboration among cities and universities involved in performance assessment activities should be encouraged as an efficient means of trying out schemas, and refining scoring protocols and technical methodologies.

It is evident then, that the demands for changes in assessment are not independent of the broader demand for reform and restructuring. For a truly performance-based model to succeed, broad-reaching changes must occur not just in instrumentation, but in curricula and in reflective and energized teaching techniques. This

reform is not about short cuts, easy answers, or simple additions to existing practices and it would be foolish to expect that such profound changes can be accomplished inexpensively and in a short period of time.

To return to my original metaphor and paraphrase a poem by Christina Rossetti entitled "Uphill":

Do the train tracks "wind up-hill all the way?

Yes, to the very end

Will the day's journey take the whole long day?

From morn to night, my friend".

References

- Ascher, C. (1990). Testing students in urban schools: current problems and new directions. *Urban Diversity Series*, No. 100. ERIC Clearinghouse on Urban Education: NY, NY. March.
- Baron, Joan Boykoff. (1990). Performance Assessment: Blurring the Edges among Assessment, Curriculum, and Instruction. In Champagne, A. B., Lovitts, B. E., & Calinger, B. J. (Ed.), *Papers from the 1990 AAAS Forum for School Science*. 127-147.
- Baron, Joan Boykoff, Ph.D. (1990). Conceptualizing and Implementing Performance Assessment on a Statewide Basis: Lessons from the Trenches. Presented at Alternatives in Statewide Educational Assessment for Early Grades: A Conference on Performance Measures, October 18-19, 1990. Wilmington, DE.
- Camp, Roberta. (April 24, 1991). Using Portfolios for Performance Assessment in Writing. *Conference on Alternative Assessment for Accountability and Instructional Improvement: Promises and Cautions*. Marlborough, Ma.
- Camp, Roberta. (1990). Thinking together about portfolios. *The Quarterly*, Spring 8-14.
- Cizek, Gregory J. (1991). Innovation or Elevation? Performance Assessment in Perspective. *Phi Delta Kappan*, May, 695-699.
- Foster, Jack D. (1990). The Role of Accountability in Kentucky's Education Reform Act of 1990. *Educational Leadership*, 34-36.
- Frechtlin & J.A. (1991). Performance assessment: Moonstruck or the real thing. *Educational Measurement - Issues and Practice*, 10(4), 23-25.
- Frederickson, Norman. (1984). The Real Test Bias. Influences of Testing on Teaching and Learning. *American Psychologist*, 39, 193-202.
- Gardner, H. (1989). Assessment in context: The alternative to standardized testing. In B. Gifford (Ed.) *Report of the Commission on Testing and Public Policy*. Boston: Kluwer Academic Publishers, Inc. 3-41.
- Hacker, Jacob and Hathaway, Walter (1991). Toward "more authentic" assessment: the big picture. Paper presented at American Educational Research Association Convention, Chicago, Illinois.
- Haney, W. & Madaus, G. (1989, May). Searching for alternatives to standardized tests: Whys, whats, and whithers. *Phi Delta Kappan*, 70, 683-687.
- Holt, Tom. (1990). *Thinking Historically: Narrative, Imagination, and Understanding*. College Entrance Examination Board. New York, New York: College Board Publications.

- LeMahieu, P. G. & Wallace, R. C. Jr. (19-). Up against the wall: psychometrics meets praxis. *Educational Measurement: Issues and Practices*, 12-16.
- Linn, Robert and Baker, Eva L. (1992). Testing as a reform tool? The CRESST Line, Winter, 1-2.
- Linn, Robert L., Baker, Eva L., and -Dunbar, Stephen B. (1991). Complex performance-based assessment: expectations and validation criteria, *Educational Researcher*, November, 15-21.
- Lockwood, Anne Tumbaugh (1991). A leap of faith. *Focus in Change*, Vol. 3, No. 1, March, 9-13.
- Lockwood, Anne Tumbaugh (1991). From telling to coaching. *Focus in Change*, Vol. 3, No". 1, March, 3-7.
- Madaus, George. (1989). On Misuse of Testing- A Conversation with George Madaus. *Educational Leadership*, May, 26-29.
- Maeroff, Gene 1. (1991). Assessing Alternative Assessment. *Phi Delta Kappan*, December, 273-281.
- Mehrans, William. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice*, Sprin& 3-20.
- Paulson, F. L., Paulson, P. R., & Meyer, C. A. (1991). What Makes a Portfolio a Portfolio? *Educational Leadership*, February, 60-63.
- Resnick, L. B. & Resnick, D. P. (19-). *Assessing the Thinking Curriculum: New Tools for Educational Reform*. 77-132.
- Rosetti, Christina G. (1966). "Uphill". *The Oxford Dictionary of Quotations*, 410.
- Simmons, Jay. (1990). Portfolios as Large-scale Assessment. *Language Arts*, 67, No. 3, March, 263-268.
- Stiggins, R. J. (1987). Design and Development of Performance Assessments. *Educational Measurement: Issues and Practices*, Fall, 33-42.
- Wiggins, Grant (1987). Design and development of performance assessments. *Educational Measurement. Issues and Practice*. Fall, 33-42.
- Wiggins, Grant (1989). Questions and answers on authentic assessment. Paper presented at Curriculum/Assessment Alignment Conference.
- Wiggins, Grant (1989). A true test: toward more authentic and equitable assessment. *Phi Delta Kappan*, 70, 703-713.
- Wolf, D.P. (1989). Portfolio assessment: sampling student work. *Educational Leadership*, 46(7) 35-39.

Discussion I

Judy Arter

Northwest Regional Education Laboratory

I will begin by discussing each of the papers individually, and then try to tie them together and relate them to the topic of this symposium.

Individual Papers

Maryellen Donahue's paper provides a good outline of various issues confronting districts pursuing performance assessment. She makes good points concerning the need to determine purpose for assessment first, the need to experiment and develop performance assessments in a low stakes environment, the need for teacher involvement and training and the need to progress slowly if our change in assessment techniques is really going to have any lasting positive impact. I found the paper to be very general however. I kept wanting to have the author give me specific ideas concerning the points she was making - for example, what strategies they used to keep the process from becoming too political, or how they kept their project low stakes.

Paul LeMahieu's paper will be very useful to practitioners. The paper is essentially a series of handouts and working documents developed as part of the Propel project, which includes a portfolio system under development in the Pittsburgh public schools since 1987. The handouts include a brief description of the portfolio system, and describes Propel's new initiative to provide large-scale accountability information by combining information across individual student portfolios. Specifically, the paper provides many sample forms and outlines that students and teachers use when assembling student portfolios, and that parents use when reviewing the portfolios. Additionally, the author includes an outline of the steps Propel went through when developing performance criteria for student portfolios, and includes the performance criteria themselves.

LeMahieu stresses that assessment strategies can serve both monitoring and instructional functions and that Pittsburgh Public Schools is beginning to believe that, if done carefully, large-scale accountability can result from a system originally developed for instruction. As one moves further away from the classroom data can and should be reduced and summarized. The problem with past assessment is that the data is reduced immediately and this is not helpful to teachers. The author refers to the procedure they developed for aggregating information across students "good faith accounting." Performance criteria are applied systematically to student portfolios, and then "external auditors" rescore a sample of the portfolios to validate the ratings.

The materials will be most useful to those who already know what they want in their portfolio system and are looking for additional ideas or ways to do things because little or no rationale or help with use is provided for the various forms included. (There is, however, rationale provided for the aggregation portion.) Additionally, it will be

interesting to see how the "external audit" procedure works. In order to rate the portfolios readers need to be trained in the criteria to achieve consistency. This might produce something of a paradox if you train for consistency how "external" can the audit be? If you don't train for consistency the external auditors might not have the expertise necessary to do the rating.

Roberta Camp's presentation was without a paper, although she did supply me with a partial draft of a paper that will form a chapter of a book. Her presentation was rich with samples of student work used to make some good points: the power of doing portfolios lies more in the process (which create dialogues between students and teachers) than in the product; it is possible to aggregate and summarize across individual student portfolios; and it is possible to build criteria for such summaries.

Robert Zlomek's presentation was also without a paper. He briefly described some of the assessment development efforts and issues in the Cedar Rapids School District.

Analysis and Summary of Papers

The papers in this session did not discuss traditional technical issues surrounding alternative assessment (accurate measures of student skills, etc.). They emphasized consequential validity (Linn et. al, 1991) how the assessment procedures and design will affect instruction, student self-evaluation, policy, etc.

The papers discussed all the following consequences:

1. Assessment is a communication device. As LeMahieu states in his paper "Methods of student assessment deliver influential messages, not only about what is valued as educational outcomes, but the educational experiences and means of teaching that are most appropriate." Given this, we have to ensure that we are sending the right pedagogical messages. The LeMahieu and Camp presentations especially focused on sending the correct messages.
2. Performance assessments can be powerful instructional tools if approached in the proper way. Performance assessment gives us the opportunity to, clarify our targets for students by developing clearly thought out and detailed performance criteria. This has the most power not for large-scale summary of student skills levels, but in clarifying targets for teacher day to day instruction of students, and student self-evaluation.

This point is expressed in various ways by the presenters: Teachers and students learn about the "standards of good work with respect to more valued outcomes" (LeMahieu). Through self reflection and with support students "gradually acquire the language necessary to describe what they see in their work...this language can evolve into personal criteria and standards for performance" (Camp).

If we don't design performance assessments do facilitate and encourage teacher and student self-reflection, then we have lost the real value of doing it.

3. All four presenters emphasized the need to proceed slowly if we are to be successful. First the discussion between teachers, students, parents. etc, is the valuable part of doing performance assessment and this requires time to train and reflect on what we value and how what we value looks in the dally performance of students. Second, it takes time to change thinking patterns, beliefs and actions. Real change takes time. The portfolio project in Pittsburgh has been evolving since 1987, and the efforts in Cedar Rapids have been in progress for over four years.
4. Although it is possible to use portfolios for public accounting, the most useful way to proceed is to develop the system for instructional usefulness first, and then think of ways to aggregate it upwards. After all, the real purpose of assessment is to improve instruction.
5. Political rather than pedagogical reasons for assessment cause some of the unfortunate consequential side effects. For example:
 - a. The pressure to go too fast. Sometimes it seems that politicians assume that we already know how to do the things they decide we should do, and that all we need is to have someone tell us to do them. The actuality is that real change takes time.
 - b. Many of the assessment systems established by the political process are high stakes in nature. The unintended side effect of this is to encourage us to want "look good" rather than to have an honest appraisal of where we are.
 - c. Some decisions are based on political expedience and not what is pedagogically sound. Some of the assessment scenarios described in this session and others at AREA this year had nothing to do with what made sense instructionally. Rather they had to do with the economy and the resulting pressure on Congress to do something, who gets elected, which special interest groups become powerful, what programs get funded, etc..

Given the negative consequential effects that the political process can have on assessment maybe we should proceed, as the presenters in this sessions have done, to design assessment systems that have a positive effect on instruction and students, and worry about accountability later.

Discussion II

Peter Volmut

Multnomah (OR) RSD

Since some of the panelist summaries weren't available and knowing Arter's competency in this arena, I felt I would take the liberty of posing two issues related to the larger context within which the alternative assessment movement is taking place.

Issue 1

Authenticity/Reality within school

Frechtling (1991) has suggested one needs to know what aspects of traditional testing-are "the problem" so that the "new" methodology can be used specifically to fix that problem--and, I would add, to avoid creating unnecessary new problem(s)! For example, it is easy to follow the logic of LeMahieu's statement that reductionism of data under machine-scored tests was the reason for Pittsburgh's moving to a system of writing portfolios.

But what happens when the connotations of a word serve themselves as the rationale for change? Authentic comes to mind these days. Meyer (1992) carefully defines authentic assessment as a subset of performance assessment. One of the key characteristics she (and others) mention is that the student's performance is defined within a "real life" context. She means by this that the locus of control rests with the student.

A decade ago when consumer movement support made Personal Finance all the rage, students spent much time being assessed via such exercises as dividing the 72 cent price by the 6 ounce weight of one can of tuna and the 67 cent price by the 5.5 ounce weight of a second can of tuna to determine that the former was a better buy per ounce. But a sage curriculum colleague pointed out that once the youngsters left the school (**but remained possessors of the locus of control**) they would buy Star-Kist because of Charlie the Tuna or the blue can because their family had always bought the blue can. School, he said, is but a simulation of real life! In other words, it may well be that the boundaries of the locus of control for scholastic and secular purposes are just not the same.

From another perspective, please note that predominant descriptions of performance assessment communication are in terms of reading and writing, even while America's non-school real-life predominant communication modes are watching and speaking! It has been thirteen years since Neil Postman (1979) described television as a curriculum antithetical to that of the schools and seven (1985) since he re-emphasized that the TV curriculum's instructional modality was entertainment--postulating a new concept that teaching and entertainment are inseparable. Yet, we proceed as though what he has written has no relevance.

Issue 2

Will the real assessment please stand up!

The 1991 Oregon Legislature passed, with little debate, a law designed to elevate the state to a position of prominence in the country by the year 2000. People who heard about Oregon House Bill 3565 felt it was either outstanding or that heavy rain had created a coating of mildew on our brains. The law is a mixture of Kentucky reform, District of Columbia business testimony, and its author's educational philosophy, moving Oregon into a more early childhood/less high school model. Related to this symposium, by 1996 all students will undergo performance assessments to determine their initial and advanced mastery certifications.

Less well known is that a second Oregon law created a Workforce Quality Council which, among its defined powers, will have a write-off on every school district budget in the state. Four--and only four--of the 26 legally enunciated members of the Council are educators. Three are the State Superintendent of Public Instruction, Commissioner of Community Colleges, and Chancellor of Higher Education, leaving one individual to represent all other educators.

The "Vision" of the council is that Oregon will have the best educated and prepared workforce in America by 2000 and in the world by 2010, involving business and labor in the development of education and training policies and providing subsequent programs for students and workers under world class standards of achievement.

Now let me present to you part of a proposed educational compact designed to allow us to win the economic battle we are engaged in with Japan and the European Community. Among the compact's principles:

- We need to cease worrying about the top 25% of our students. We have done a good job with them and will continue to do so. We need to concentrate our efforts on the bottom 50% of the student population.
- Business will have to have workers who can solve Algebra problems so that in turn they can do the Operations Research work required by their process-oriented jobs. Obtaining right answers will be at least on a par with learning to learn and self-awareness.
- Because of this business need, each state's business community should write an achievement test which would cover what the high school graduate needs to know to work in American firms. Students who passed such a test would have it noted and firms would commit to hire them. "It wouldn't be a test written by ivory tower professors or educational bureaucrats!"

When I read this proposed compact, the Oregon system became much clearer to me. Where the (authentic?) performance assessment train is going while business communities write the (pragmatic?) tests of this proposal is an issue that should be of concern to us. The compact is proposed by Lester Thurow (1992), dean of MIT's School of Management and one of America's most respected economists.

Perhaps the irony of this session may be that the engineer of Donahue's train is in Cambridge, just across the river from her Boston office!

References

Frechtling, J (1991). Performance Assessment: Moonstruck or the Real Thing? Educational Measurement: Issues and Practice, 10(4), 23-5.

Meyer, C. A. (1992). What's in a Label? Performance Assessment or Authentic Assessment. Educational Leadership, 49(8),.

Postman, N. (1979). The First Curriculum: Comparing School and Television. Kappan, 1.1(3), 163-8.

Postman, N. (1985). Amusing Ourselves to Death. New York: Viking Penguin, Inc.

Thurow, L. (1992). Head to Head. New York: William Morrow and Co.