

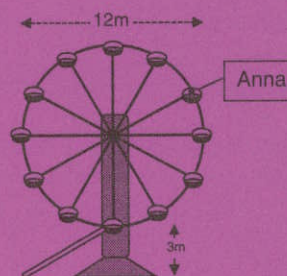
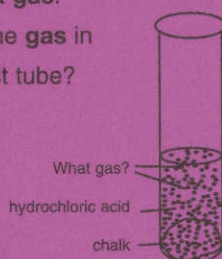


*Chalk is added to a test tube containing hydrochloric acid. A gas evolves. Identify the gas that evolves.*

Jane puts some **hydrochloric acid** and some **chalk** into a test tube.

It makes a **gas**.

**What is the gas** in Jane's test tube?



How high off the ground is Anna?

# An Approach to Test Development

Tandi Clausen-May



# **An Approach to Test Development**

Tandi Clausen-May

Published in January 2001  
by the National Foundation for Educational Research,  
The Mere, Upton Park, Slough, Berkshire SL1 2DQ

© National Foundation for Educational Research 2001

Registered charity No. 313392  
ISBN 0 7005 3021 5

# **Contents**

<b>Foreword</b>	<b>i</b>
<b>Introduction</b>	<b>1</b>
<b>1 Some Basic Issues</b>	
1a) Two key ideas: validity and reliability	4
1b) The impact of assessment	7
<b>2 Developing Tests</b>	
2a) A holistic model	13
2b) Coverage	14
2c) Exploiting good question contexts	17
<b>3 Writing Questions</b>	
3a) Types of question	23
3b) Contexts	29
3c) Differential item functioning	31
<b>4 Accessibility</b>	
4a) Should tests be accessible?	35
4b) Language and sentence structure	38
4c) Test layout	42
4d) Modifications and special arrangements	45
<b>5 The Test Development Cycle</b>	
5a) Organising question writing	48
5b) Informal trials	50
5c) Formal trials	53
5d) Developing robust mark schemes	57
5e) Cycles and spirals	64
<b>6 Statistics for Test Users</b>	
6a) Another look at validity	66
6b) Reliability as a statistical concept	70
6c) Standardised scores	72
6d) Measurement error, true scores and confidence bands	75
6e) Cut scores	77

<b>7</b>	<b>Statistics for Test Developers</b>	
7a)	Samples	81
7b)	Facilities	82
7c)	Point biserial correlation coefficients	85
<b>8</b>	<b>Looking Ahead</b>	
8a)	Information and communications technology	90
8b)	The beginnings of change	92
8c)	What is to come?	95
8d)	Computer-adaptive testing	99
	<b>Appendix</b>	
	Writing multiple-choice and multiple-response questions	108
	<b>References</b>	112

## **Foreword**

There is no shortage of academic books and journal articles about assessment and testing – new contributions appear with increasing frequency. However, this reader often wishes that some of those writing could show evidence of, and take account of, practical experience of setting and refining test questions.

Those who have such experience tend to avoid writing about it. They are too busy with their teaching or their research when the stresses and urgencies of examination times have passed. Yet both they and users of test and test results could study with profit some of the technical and professional guidelines that could enrich their understanding of test issues.

The author has judged well in limiting the scope, and therefore the length and complexity, of this work. Because of these decisions, the writing can achieve both clarity and brevity, and this book can therefore be welcomed, both for the purposes it serves and for the way in which it serves these purposes.

Paul Black  
Emeritus Professor of Science Education  
King's College, London



## **Introduction**

Like it or not, our education system is becoming more and more test-driven. A glance at the Education section of any bookshop will reveal shelves full of published tests, pitched at a range of levels, covering a variety of subjects. And that is before you even start to look through publishers' catalogues, or take account of official assessment structures – Key Stage tests, GCSEs, A-levels, NVQs, Key Skills tests, and so forth – imposed on, or offered to, schools and colleges.

As access to education increases world-wide, the use of formal tests is likely to increase rather than decrease. Decisions relating to selection and promotion must be made somehow, and tests can offer at least a degree of fairness. They make it possible – not certain, perhaps, but possible – for selection and recruitment to be based on factors other than wealth, political position or family background. Formal tests may not always be fair, but they are less obviously unfair than selection based on influence.

With the rapid increase in the number of tests which are available, it is inevitable that some will be better than others. So the first group of people to whom this book is addressed are those who use paper-and-pencil tests, and particularly those who have to choose tests for others to take. There are good tests, and there are some that are not so good. The aim is to help users to make the best choice they can out of the available options.

As formal tests are used more and more, an increasing number of people find themselves involved with writing or reviewing such tests as part of their job. Teachers, of course, have always written end-of-term, end-of-year, end-of-unit, or whatever, tests. But even teachers may be writing more tests than they used to. And formal tests in other contexts – Key Skills tests for people training to be anything from stable hands to nursery nurses; job selection tests, for any job you can imagine; written driving tests, for anyone who wants to take a car or a lorry on the road – all



these tests have to be developed and reviewed. The people who are involved in the development of these new tests should know something about the subjects being assessed, and should have worked with the learners who will be taking them. This new group of question writers and reviewers may not be steeped in educational philosophy and psychology: their experience and expertise lie in other areas. A second aim of this book is to help new test developers to understand some of the issues, and to recognise some of the pitfalls, which surround the creation of a formal test.

A book of this size cannot cover everything for everyone, but it can consider at least some of the problems which the developers of a variety of different types of formal test are likely to meet. The examples of questions and mark schemes given to demonstrate the process of test development are for the most part taken from the primary mathematics and science curriculum. The assessment of other subject areas, such as English or the humanities, or of more practical activities, involves a number of wider considerations, many of which are not covered here. Projects, coursework, and other more extended pieces of work which can form the basis of an assessment are not discussed. None the less, much of what is presented here may be relevant to a range of tests in different subject areas.

The book starts with a discussion of *Some Basic Issues* in Chapter 1. It goes on to examine aspects of the process of *Developing Tests* in Chapter 2, offering more detailed advice in relation to *Writing Questions* in Chapter 3, with a particular focus on the issue of *Accessibility* in Chapter 4. The organisation of the whole *Test Development Cycle* is then considered in Chapter 5, with a brief introduction to some aspects of *Statistics for Test Users* in Chapter 6, and *Statistics for Test Developers* in Chapter 7. Finally, Chapter 8 involves *Looking Ahead* at ways in which the development of information

and communications technology may influence tests in the future.

An indication is given at the beginning of each chapter of the main points covered, and of the aspects which may be of particular interest to different readers. Test users are likely to find the whole of Chapters 1, 4, 6, and perhaps 8, and the first part of Chapter 2, of particular interest. Chapters 1 to 4, 5, 6, and possibly 8 may offer useful guidance to question writers and reviewers, while Chapter 7 may be of greater interest to those with overall responsibility for the development of a formal test or examination.

### ***Introduction: Key Points***

Formal tests may not always be fair, but they are less obviously unfair than some other forms of selection.

This book is written to help:

- those who use tests for different purposes;
- those who contribute to the development of formal tests, for example as question writers or reviewers;
- those who have overall responsibility for the development of a test or an examination.

## 1 Some Basic Issues

*This chapter starts with a discussion of two key ideas which are basic to the development of any test: **validity** and **reliability**. Test developers constantly strive to achieve both, but can never be certain that they have achieved either. The chapter goes on to consider the impact which the assessment structure is likely to have on the curriculum, and to discuss ways in which a test may support, or may tend to undermine, teaching and learning in the classroom. Both sections should be of interest to anyone concerned with the development or use of formal tests.*

### ***1a) Two key ideas: validity and reliability***

***Validity:*** *The test tests what it claims to test.*

But can we ever be sure that a test really does test what it claims to test? To take an obvious example: if I took a driving theory test designed for ordinary drivers, then I ought to pass it. If I did not, then that would say something about my driving – or at least about my knowledge of driving theory. But supposing the test were given to me in Russian? In that case, of course, I would expect to fail. But I would have 'failed' the test situation, not the test subject. My failure might be valid evidence of my knowledge of Russian; it would not be valid evidence of my knowledge of driving theory.

Again, formal tests like the ones discussed here are relatively easy to manage in the classroom – but some things are much easier than others to test with pencil and paper. For example, is a driving theory test, in any language, a valid test to give to would-be drivers? Is a trainee gardener's knowledge of the chemical composition of fertiliser, or her ability to work out the area of a circle to two decimal places, a valid assessment of her ability to choose a fertiliser and decide how much to spread on a rose bed? A real rose bed is likely to be an irregular shape, and

it is most unlikely to be a perfect circle – but a paper-and-pencil test of numeracy, for example, will focus on the theoretical mathematics, because, unlike fertiliser spreading, it is something that can be assessed easily in examination conditions. Many formal written tests are probably not valid assessments of the ability to actually do a real job – but we often assume that they are, because this form of assessment is more manageable.

Even if we overcome our qualms about the validity of the content of the test, other issues relating to its accessibility must be considered. I know that I have no knowledge of Russian, so I would not expect to be able to pass a Russian driving theory test – and, indeed, I would make a great deal of fuss if anybody tried to make me take one. But the situation is not always so straightforward. Many aspects of the language, layout or presentation of a question paper may affect its validity. It may use language which is unfamiliar, or contexts which are confusing. Similarly, the test situation itself may constitute a barrier to assessment. If the person taking the test is so anxious that they cannot function normally, then the test is at least in part a measure of their ability to control their panic, rather than to answer the questions. A lot of effort may go in to making the test as accessible as possible, in order to ensure that it is actually assessing what it purports to assess – but no one can ever know, for each and every candidate, just what there is in the test situation which might affect their performance.

So we can never be certain how valid a test is. Some alternative ways to define validity are discussed later, in Chapter 6, but, for the moment, it is enough to remember that we can never be sure that a test actually measures the abilities that it claims to measure, and nothing else, for the person taking it.

So much for validity. What about reliability? Can we be sure that the test is reliable?

***Reliability:*** *If the same person had taken the test for the first time on another occasion, then they would have got the same result.*

But that is clearly absurd. The same person cannot take the same test for the **first** time on another occasion. As Ian Schagen explains in his article on the statistics of tests, 'Testing, testing, testing',

In theory, in order to measure reliability we would need to 'brain-wipe' a set of candidates and make them do the test again, with no memories of questions or answers from their previous attempt, or tiredness or change of mood. Impossible, of course.

(Schagen, 1999, pp. 28-9)

If a group of candidates do take the test twice, then one of the occasions must be the second time. Normally, people do a bit better the second time around – although when this sort of experiment is carried out, there are usually a few who actually do less well the second time. But in any case, asking the same people to take the same test on two different occasions does not give a true measure of the test's reliability. For that we would somehow have to wipe the slate clean, erasing their memory of their first experience of taking the test before they took it again. So in practice the *reliability* of a test may be a useful statistically defined concept, but it can never be proven.

These two key ideas, validity and reliability, constitute the Holy Grail of test development. We constantly seek them, and we go to great pains to try to secure them, but we can never, ever, be certain that we have achieved them.

Furthermore, the two key ideas are in conflict. To make a test valid we must do our best to ensure that it is accessible – that the people taking the test are actually responding to the questions asked, and not to some other, extraneous factors. But this

requires a degree of flexibility – offering another opportunity to do the test if the person taking it is tired or upset; reading out the questions, and perhaps rewording them if the language is unfamiliar; giving the test taker more time if they are slow readers or writers. But varying the conditions under which the test is taken in order to ensure its *validity* may make it even less likely that a person taking the test would get the same result the second time around – so it may tend to undermine the *reliability* of the test. If the same person took the same test for the first time on another occasion, then they might, perhaps, get pretty much the same result – but if the conditions under which they took it changed, then this would be less likely. Thus there is always a tension between ensuring that a test is accessible to a range of users, to make it valid, and always administering it in exactly the same way to everybody under 'controlled' conditions, to improve its reliability.

So we have two aims – to make tests valid, and to make them reliable – but there is often a tension between them. We can never be sure that we have achieved either aim, and any attempt to ensure one may tend to undermine the other. Our task as test developers is to reach a compromise: to develop an assessment which is as valid and as reliable as possible, but not to forget the inevitable limitations of any test we can produce.

### ***1b) The impact of assessment***

As every teacher knows, *What You Test Is What You Teach*. If a particular topic is in the curriculum, but the pupils' knowledge and understanding of that topic are never assessed, then quite soon it is likely to be marginalised and to receive less emphasis in the classroom. But equally, if an area of the curriculum is to be assessed for the first time, then a lot of effort will go into the teaching and learning of the new material. This has been evident recently in a number of contexts. For example, the introduction

in 1998 of the mental tests as part of the national Key Stage 2 and 3 assessments in mathematics resulted in a rapid increase in the time spent in primary and secondary school classrooms on the development of the relevant mental skills. Similarly, the introduction of tests in Numeracy, Communication, and IT for students undertaking initial teacher training is designed to increase the time that they spend gaining experience of all three Key Skills.

So in any curriculum area, what is actually taught and learnt is likely to be influenced strongly by the tests which are used to assess achievement. Teachers may use the assessments to decide what they should teach and how they should approach each topic. This can happen whether or not the assessments were actually intended to be formative – to provide the teacher with information on which to base their decisions. If the questions in a test encourage a 'mechanistic' approach, with answers that can be learnt by rote and the routine application of standard algorithms to easily predictable types of question, then, in some cases at least, that is likely to be what is taught. Only if the questions assess the pupils' understanding of the principles which underlie their factual knowledge and the methods they use will this be the focus of teaching and learning in the classroom.

But it is not easy to develop questions which can distinguish between pupils who have a sound *conceptual* understanding of the principles which underlie the methods they use, and those who have only *procedural* knowledge which may be based largely on rote learning. Questions which assess understanding are likely to be quite long, with a lot of information to be gathered off the page. They may require an open response, which is often hard to mark. For example, the 1997 Key Stage 2 science test included a question about the shadow which was formed on a screen when a torch was shone on to a puppet. The pupils were asked to:

Explain how a shadow is formed.

(1997 Key Stage 2 Science Test A)

The mark scheme required pupils to demonstrate that they understood the principles involved by showing 'an awareness that a shadow is formed when light is blocked by an object'. The examples of acceptable responses indicated that *the light (or the sun) is blocked/stopped/can't pass through* was worth the mark. On the other hand, *a shadow is a place where there is no/little light* was 'not an explanation of how a shadow is formed', so it was ruled out by the mark scheme. But consider the following responses:

*The shadow is in the place where the light cannot pass through to get to the screen.*

*The shadow is in the place where the light cannot get through to the screen.*

*The shadow is in the place where there is no light getting through to the screen.*

*The shadow is in the place where there is no light getting to the screen.*

*The shadow is in the place where there is no light on the screen.*

The first of these responses is clearly acceptable according to the mark scheme, and the last is unacceptable. A judgement would have to be made, however, as to where to draw the line for the responses in between.

To avoid some of the problems which can arise when responses to an open question like those given above have to be marked, the attempt may be made to assess pupils' understanding through a more closed question. To take one example, with which many mathematics teachers have struggled over time: the 'Four Rules for Fractions' are often taught with a drill and practice approach, because they are so amenable to the simple repetition of the 'rules'. So pupils may acquire a set of instructions for the

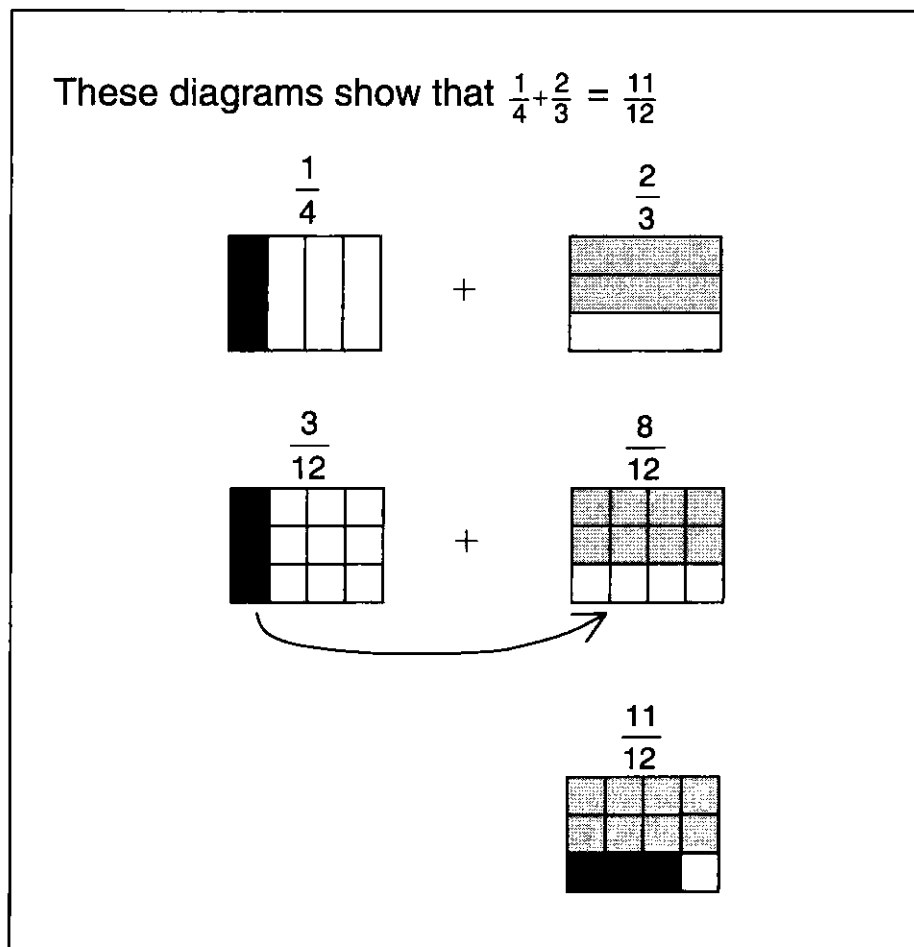


addition, subtraction, multiplication and division of fractions which have very little meaning – and which, as adults if not before, they rapidly forget. A test question such as

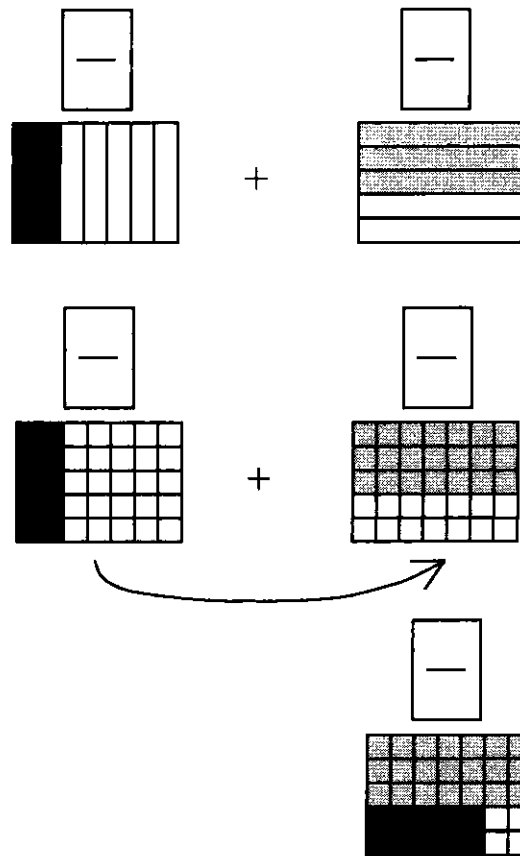
$$\frac{2}{7} + \frac{3}{5} =$$

will tend to encourage this approach, since pupils who can remember the rules for the addition of fractions are likely to get the mark, whether or not they understand why the rules work.

To get away from the routine application of meaningless algorithms, a question was developed which focused on the pupils' understanding of the principles which underlie the rules for the addition of fractions.



Fill in the boxes to say what these diagrams show:



This question goes some way towards assessing the pupils' understanding of the meaning behind the process of finding a common denominator when two fractions are added.

But the question occupies a whole page of the test, in contrast to the simple ' $\frac{2}{7} + \frac{3}{5} =$ '. It requires the pupil to study and understand the example given, and offers no credit for the routine application of the rules for the addition of fractions. None the less, some commentators saw the question as 'helping' pupils to carry out a computation which they should know how to do

anyway. It was seen as a good teaching technique, but inappropriate for an assessment. It was 'unfair', since a pupil who did the addition by another method (for example, by following the rules) might not understand the diagrams and might fail to fill in all the boxes correctly. It was argued that, although the expected answers are  $\frac{2}{7}, \frac{3}{5}, \frac{10}{35}, \frac{21}{35}$  and  $\frac{31}{35}$ , so long as the first and last line had been completed correctly the pupil had done the computation and should be awarded the marks. Altogether the question was regarded as being much too complicated, when the simple, straightforward ' $\frac{2}{7} + \frac{3}{5} =$ ' would serve just as well. This attempt to develop a closed question which would assess the pupils' conceptual understanding was rejected.

So it may be possible to develop questions which focus on the pupils' understanding of mathematical or scientific concepts – but it is certainly not easy. Successful questions are often open-ended, so they require a mark scheme which takes account of the pupils' understanding however this may be expressed. Both the questions and the mark scheme are much more difficult, and time consuming, to develop than a straightforward closed-response or multiple-choice question.

### ***Some Basic Issues: Key Points***

*Validity* and *reliability* constitute the 'Holy Grail' of test development. To be manageable, any test has to be a compromise with regard to both.

Only if the questions in a test assess the pupils' understanding will this be the focus of teaching and learning in the classroom.

Questions which assess understanding are likely to be more complex and harder to mark.

## 2 Developing Tests

*The way in which the development of a test is planned and the questions are written will have a significant effect on the nature of the resulting test. In this chapter, a holistic approach is discussed, and it is suggested that questions may be developed which seek to bring out the links between different aspects of the subject being assessed. The first section of the chapter is likely to be of general interest. The second and third, which focus more closely on the detail of question development, may be more relevant to question writers and reviewers, as well as to those with overall responsibility for the development of a test.*

### **2a) A holistic model**

Assessment may be approached in a variety of different ways, from open-ended tasks through to timed, pencil-and-paper, machine-markable tests. Again, a test may be narrowly focused on a specific aspect of the curriculum, or it may be wide-ranging, covering a variety of different topics at different levels. But regardless of the nature of the questions, and of the range of topics covered, some tests have a much more coherent feel than others. The test as a whole may serve to emphasise the connections between different aspects of the subject, and to bring out the common structures which underlie the different topics covered. Alternatively, the test can be composed of a series of atomised, disconnected questions, which give no feeling for the relevance of one part of the curriculum to any other. So, for example, a primary science test might have different sections covering aspects of *Life processes and living things*, *Materials and their properties*, and *Physical processes* (or biology, chemistry and physics): this would tend to encourage the teaching of these areas as separate, isolated subjects. On the other hand, a question on, say, *Water* might cover a range of topics drawn from different areas of the science

curriculum. If tables and graphs showing the results of some observations were presented in the question, then aspects of the mathematics curriculum might also be covered. Such a question would go beyond the testing of individual little segments of knowledge, and allow for the assessment of the pupils' understanding of the interrelationships which form the basis of a sound understanding of science and mathematics. It would also help to encourage an approach in the classroom which emphasised the connections between these areas of the curriculum.

The way in which the development process is planned and carried out can have a profound effect on the overall quality of the final test. Questions which relate clearly to the pupils' own interests will go a long way to make the test relevant and motivating. A series of questions leading from one scenario is likely to be more meaningful than a collection of unrelated 'items'. Planning the test as a whole, rather than in disconnected bits, will help the developer to achieve the goal of a coherent assessment.

## ***2b) Coverage***

When a new test is going to be developed, the first task is to decide on its content – what it is that the test is supposed to be testing. This is often quite formally specified in commercially produced or national tests. But in every list of topics to be covered there will always be some areas on which it is much easier to set test questions than others. The danger is that a test will cover only what is easily, and relatively reliably, testable. But since *What You Test Is What You Teach*, this can have an invidious effect on the curriculum. This being so, test developers are bound to at least attempt to assess some topics which are not straightforward. These questions are necessary to ensure that the coverage of the test is balanced.

At first glance, the specification for a test may look terribly restrictive. For example, the 'Number' section of a mental mathematics test for nine-year-olds had the following extremely detailed specification laid down:

	In context	Either	Not in context	Total
Addition	1		1	2
Subtraction	1		1	2
Multiplication	1	1	1	3
Division	1	1	1	3
Place Value		1		1
Neg. Nos.		1		1
Fractions		1		1
Decimals		1		1
Total	7		7	14

This coverage chart indicates that *Addition* and *Subtraction* should each have exactly two marks, one in and one out of context, while *Multiplication* and *Division* should each have three, with at least one in and one out of context. Each of the other topics listed were allowed only one mark – but *Estimation*, for example, was not mentioned. This sort of list is inevitably arbitrary and selective, and could be very constraining.

Fortunately, however, there is actually far more flexibility than may at first appear. Few good test questions fall unquestionably into only one category: most could be slotted into different positions on the coverage chart, and could be 'counted' for a

number of different topics. For example, consider the following multiple-choice question:

$$397 + 456 = 853$$

The number fact in the box is true.

Which one of the following number facts is true?

- a)  $497 + 466 = 863$
- b)  $853 - 456 = 403$
- c)  $39.7 + 45.6 = 8.53$
- d)  $387 + 446 = 833$

This is an uncontextualised question which requires pupils to work from a given addition fact – that  $397 + 456 = 853$  – to check some other possible number facts.

The question clearly relates to the addition of two three-digit numbers – although pupils are not required to actually add the two numbers, as this is done for them. The first option requires them to consider the effect of adding two numbers, one of which is 100 more, and one 10 more, than the originals, and to compare the result with the original total. Pupils may respond by adding 110 to 853, or by subtracting 110 from 863. The second option is a subtraction, and explores this operation as the inverse of addition. Option c) relates to the addition of decimal numbers, and focuses on place value. A good approach here would be to use estimation to realise that 'about 40' plus 'about 45' must be 'about 85', not 'about 8.5'. Finally, option d) (which is the only

correct option) again focuses on place value: the pupil must realise that each of the two numbers to be added is 10 less than the originals, so the total must be 20 less.

Thus this question involves *Addition*, *Subtraction*, *Place Value* and *Decimals*, and also possibly *Estimation*, which was not one of the topics listed in the coverage chart for this particular test. The degree to which questions can be slotted in to different positions on the coverage chart will vary, but in this case the question could be attributed to any of four out of the eight Number topics covered by the test. This can give the test developer a lot of flexibility – although there will always be some topics for which it seems almost impossible to find good test questions.

Although it is tedious, it is worth keeping track of all the topics covered by each mark as the questions are written. This will allow the best use to be made of the flexibility which multiple attributions offer when the test paper is finally put together to have exactly two marks for *Addition*, and only one mark for *Place Value*, or whatever. The use of a spreadsheet will reduce the need for repetitive computations in a test with a lot of marks.

### ***2c) Exploiting good question contexts***

We have seen how a good question may cover a wide range of topics, and so may be attributed to a variety of different skills. Given an 'itemised specification', with exactly one mark for this and two for that, there is a great temptation to start at the beginning and try to think of a question for each topic to be covered in turn. But this tends to lead to a very atomised test, full of discrete, disconnected questions. Here again, since *What You Test Is What You Teach*, this may have a bad effect on the curriculum. If questions 1 and 2 are on addition, questions 3 and 4 on subtraction, questions 5, 6 and 7 on multiplication, and so



on, then eventually, as the curriculum backwash takes effect, week 1 may be spent doing additions, week 2 doing subtractions, week 3 multiplying, and so forth.

A better approach to test construction is to start with the questions rather than the topics. Find rich question contexts, develop them in any direction they can go, and then see which of the topics mentioned in the specification have been covered. This holistic test development model will produce a much more coherent test, with room for some more searching questions, and ideas which can be taken up and extended later in the classroom.

This approach can be taken even within the restrictions of multiple-choice, machine-markable questions. For example, the question below was developed for use in a test of the *Application of Number*, one of the Key Skills associated with a wide range of National Vocational Qualifications:

The table shows the times taken by five runners in a race:

Andress	Battula	Collins	Derrigo	Evans
3 min 13 sec	2 min 56 sec	3 min 29 sec	3 min 7 sec	2 min 49 sec

How many seconds behind the winner was the runner who came third?

a) 16 secs    b) 18 secs    c) 22 secs    d) 58 secs

This question requires candidates to use a range of skills which are specified in the syllabus for *Application of Number* at level 1, including the ability to *read and understand straightforward*

*tables, to read and understand numbers used in different ways, and to identify suitable calculations to get the results you need.* It could therefore be slotted into a number of different positions on the coverage chart, and is the sort of question which gives the test developer considerable freedom from the strait-jacket of a rigid specification.

The context of the times taken by runners in a race could be exploited further with questions such as the following:

What was the range in the times taken?

- a) 24 secs    b) 40 secs    c) 49 secs    d) 80 secs

What was Battula's time rounded to the nearest 10 seconds?

- a) 2 mins  
b) 2 mins 50 secs  
c) 2 mins 55 secs  
d) 3 mins

These questions address two more of the skills which are specified in the syllabus for *Application of Number* at level 1: the ability to *find the range for up to 10 items*, and to *work to the level of accuracy you have been told to use*. Thus the series of questions, all leading from the same scenario, ranges over different aspects of the syllabus to be covered, and encourages a more holistic approach in the classroom. Admittedly a few 'gap-filling' questions may still be inevitable, but experienced question writers will be aware of where the gaps in coverage are

likely to occur, and will keep a special look-out for opportunities to fill them within the context of any question being developed.

When a series of questions leading from a common scenario is written, however, care must be taken to avoid the effects of 'follow-through'. For example, the question *Who is the winner?* could be asked of the race data, followed by the question *Who came last?* But it could be argued that candidates who made the common error of taking the highest number on the table to be the winner might, on follow-through, be awarded the second mark if they took the lowest number in the table to be the loser. Otherwise, the two questions would penalise the candidate for the same mistake.

Again, consider the following question:

Peter buys four sandwiches at 75p each.

- a) How much must he pay?
- b) Peter pays with a five pound note.  
How much change will he get?

The correct answers are, of course, £3 and £2. A pupil who calculated the answer to part a) correctly, but then made an error with part b), would get credit for their first answer, but lose the second mark. But what about a pupil who made a mistake with the first part, but followed through correctly in the second? Supposing, for example, that they wrote £1.50 for part a), having allowed for only two sandwiches instead of four, but then subtracted the £1.50 correctly from £5 to get £3.50 for part b). This pupil has, if anything, carried out a more difficult calculation than the question requires in part b), correctly subtracting £1.50, rather than £3, from £5. But if follow-through is to be allowed then this must be specified in the mark scheme – and then only if it does not lower the difficulty level of

the question. Furthermore, there is always a risk that it will be missed by the marker, who is looking for the correct answer, not for an incorrect answer based on correct follow-through. If the responses are to be marked by a computer then the problem of ensuring that follow-through is recognised may be even greater.

In order to avoid the problems that can arise with follow-through marks, question writers are sometimes tempted to award only one mark for two correct answers. However, if the pupil has only a given amount of time in which to complete the test, then the developer should be aware of the average amount of time that is being allowed for each mark. For example, a test carrying 50 marks in total which must be completed in an hour gives pupils just over a minute to obtain each mark. In practice, of course, not all pupils will complete the test – so a pupil who gets only half-way through, for example, will have spent an average of over two minutes on each mark. None the less, even though pupils will actually spend different amounts of time obtaining different marks, the notion of 'one minute per mark', or whatever, is useful. If a pupil will have to spend four or five minutes drawing a graph, say, or finding information from a table or a diagram, then the question needs to carry at least three or four marks to make it worth while, and it must be clear what each mark is for. It should be possible for pupils to get some way into the question, and obtain some of the marks, without necessarily completing it correctly and getting them all. But with a holistic approach to test construction, the majority of questions will have several parts and cover a range of ideas, and will help to demonstrate the interconnected nature of the subject as a whole.

### ***Developing Tests: Key Points***

A holistic approach to test development can encourage a coherent approach in the classroom, and allow for the assessment of pupils' understanding of the interrelationships between different aspects of the subject.

Most good test questions cover a range of topics, and can be slotted into different positions on the coverage chart.

A good question context can be developed to cover different areas of the curriculum.

### 3 Writing Questions

*Some aspects of the process of question development have already been discussed in the previous chapter. Here a more detailed account is given of issues which may arise when different types of question are written. Only **multiple-choice**, **multiple-response**, **closed-answer** and **open-response** questions are considered. Examples are given which all relate to the same area of the mathematics curriculum, in order to demonstrate some of the differences – and the similarities – of these types of question. Broader, essay-type questions present some more complex problems, which are not considered here. The use of **question contexts** is also discussed, and issues relating to **differential item functioning** are considered. This chapter is likely to be of particular interest to test developers, including question writers and reviewers.*

#### **3a) Types of question**

Questions in written tests which are considered here may be broadly categorised into four types: *multiple-choice*, *multiple-response*, *closed-answer* and *open-response*. The simplest of these is probably the *multiple-choice*. This requires one clearly correct response, and at least three plausible distracters – wrong answers which do seem reasonable at first glance, but which are definitely incorrect. For example, each of the distracters in the number facts question discussed in section 2b) is the result of a plausible, but incorrect, calculation.

$$397 + 456 = 853$$

The number fact in the box is true.

Which one of the following number facts is true?

- a)  $497 + 466 = 863$
- b)  $853 - 456 = 403$
- c)  $39.7 + 45.6 = 8.53$
- d)  $387 + 446 = 833$

Option a),  $497 + 466 = 863$ , is the result a pupil would get if they increased the total by 10, to allow for the '60' in 466, but failed to increase it by a 100, to allow for the '400' in 497. A pupil selecting option b) has ignored the original addition fact, and has attempted to subtract 456 from 853, making the common error of always subtracting the smaller digit from the greater. Option c) exploits the common tendency to apply half-remembered 'rules' in inappropriate contexts, in this case the 'rule' for finding the number of digits after the decimal point in the product (not the sum) of two decimal numbers. Option d) is correct.

A good way to develop a multiple-choice question is to trial an open-response question first, asking a small group of pupils to write their own answers rather than choosing from a set of options. Then the test developer can base the question on the pupils' work, picking out common errors and incorporating them into the distracters. For example, the developer could work on the question given above with an open-response task, such as:

$$397 + 456 = 853$$

The number fact in the box is true.

You can use it to find some more number facts.  
For example:

$$497 + 466 = 963$$

$$85.3 - 45.6 = 39.7$$

Use

$$397 + 456 = 853$$

to find some more number facts.

Trialling an open-response question like this, and perhaps discussing their responses with the pupils, would help the developer to decide which sorts of error were the most common, and would thus form the best distracters.

*Multiple-response* questions look very like multiple-choice, but more than one of the options may be correct. The pupil is required to select all the correct options in order to obtain the mark. To take a simple example, a question similar to those we have already considered could be presented in multiple-response format:



$$397 + 456 = 853$$

The number fact in the box is true.

Which of the following number facts are true?

There could be more than one.

a)  $497 + 466 = 863$

b)  $852 - 456 = 396$

c)  $39.7 + 45.6 = 8.53$

d)  $387 + 446 = 833$

Here, option d) is true as before, but option b) is also true. It requires the pupil to recognise the subtraction which is the inverse of the addition given in the question, and then to subtract 1 from both the 853 and the 397.

In the multiple-response question given above, pupils were warned that there could be more than one correct response, but were not told how many they had to find. An alternative structure would ask the pupil

Which **two** of the following number facts are true?

This is slightly easier, as it limits the possibilities to be considered. Some test developers feel that pupils should always

be told how many options they should select, but there is no universal agreement on this point.

A *closed-answer* question always requires a precisely defined short answer, but there are a number of forms of presentation. For example, to take another question covering the same topic as those given above:

$$397 + 456 = 853$$

The number fact in the box is true.

Use it to work out the answer to:

$$497 + 466 =$$

Alternatively, a question might be presented in 'fill-the-gap' format:

$$397 + 456 = 853$$

The number fact in the box is true.

Use it to help you to fill in the gap:

$$397 + \underline{\hspace{2cm}} = 733$$

These examples are offered here to demonstrate some of the different ways in which what is essentially the same question may be presented. All the versions of this question – multiple-

choice, multiple-response, and closed-answer – are quite clear about what they are asking, and have precise, unambiguous answers. They are all easy to mark since no two markers could reasonably disagree about whether the pupil had got the answer right. Indeed, they could be machine marked if pupils recorded their selected responses on OMR (optical mark reader) or OCR (optical character recognition) sheets, and this would make them relatively economical to use.

But what is less certain is that these questions will distinguish between pupils who have a high level of *conceptual* understanding – who recognise the relationships between the number facts – and those who have only *procedural* knowledge – who know how to carry out computations, but have no real grasp of the principles which underlie the methods they use. Apart from a couple of traps, where distracters are offered in the multiple-choice or multiple-response versions of the question which rely on the misuse of irrelevant algorithms, there is nothing here which assesses the pupils' understanding of the techniques they may use to obtain or select the right answers. For example, in these versions of the question, the pupils are given the fact that

$$397 + 456 = 853$$

Pupils must then decide, first, whether or not  $497 + 466 = 863$ . A pupil who understands the relationships between the numbers in the given fact and those in the option will realise, without any further computation, that since one is 100 and the other is 10 more than the originals, the total will be 110 greater. But a pupil who does not have this level of understanding may simply carry out the routine addition, and reject the option this way. Either pupil will give the correct response, and we cannot tell from this whether they have shown conceptual understanding, or merely procedural knowledge.

One way to gain further insight into the pupils' level of understanding might be to scrutinise their answer sheets to see whether they had, in fact, carried out the redundant calculations. A pupil who had shown no evidence of adding 497 to 466 might be judged to have demonstrated conceptual understanding. But this change in the way that the pupils' work was marked would change the nature of the question from multiple-choice to open-response. The mark scheme would be very difficult to develop, and might well lead to inconsistencies and instances of unfair marking. For example, some pupils who had completed the multiple-choice question correctly, using the interrelationships between the numbers to decide which of the options offered were true, might then 'check' their working by carrying out the redundant calculations. This could lead the marker to judge from their written responses that the pupils had shown only procedural knowledge, rather than conceptual understanding.

The more open a question, the more difficult it is to develop a reliable mark scheme which will cover all the possible responses which pupils may make and will ensure that all markers will come to similar decisions with regard to individual pupils' answers. On the other hand, the more closed the question, the less likely it is to distinguish between pupils who understand the concept and those who merely know a method to find the answer. This is just another of the tensions which the test developer must work with, and attempt to resolve.

### ***3b) Contexts***

Putting questions into *context* can go a long way to make abstract ideas more meaningful. For example, in one test for 13-year-olds, pupils were given two questions which involved division by a fraction, one in context and one not.

In context:

How many quarter hours are there in three-and-a-half hours?

Out of context:

Divide two and a half by a quarter.

Although the computations required were very similar, only 35 per cent of the pupils were able to do the uncontextualised division, while more than twice as many – 73 per cent – answered the contextualised question correctly. There is evidence that, faced with a dry, pointless calculation, many people respond by trying to remember the correct 'rules' for the given type of computation – *I have to turn something upside down and then multiply all the numbers, don't I?* But given a calculation which has some meaning attached, they are much more likely to try to understand the situation and interpret it mathematically in order to arrive at a sensible answer, (Clausen-May, 1998).

It sometimes happens, however, that in the attempt to place everything in context, aspects of mathematics may be forced into contexts in which no one would ever actually use them. A common culprit here are questions which are designed to assess the pupil's ability to use algebraic formulae. People do understand and use given formulae quite regularly in their daily lives – so the instructions for cooking a chicken, for example, might be something like '55 minutes per kilo plus 25 minutes'. But this can become much more difficult to understand when it is translated into a mathematical formula and used in a test question:

$$m = 55k + 25$$

The algebraic formula is simply off-putting, where the instructions are clear enough for most people to understand and use.

In this case, it is probably better to use scientific or mathematical contexts, such as the formula for Ohm's Law or for the volume of a particular solid, rather than trying to make algebra more directly relevant to the pupils' everyday life. Alternatively, pupils may be assessed on their ability to follow instructions such as those for cooking a chicken, rather than on their use of inappropriate symbolically presented mathematical formulae.

### ***3c) Differential item functioning***

Another problem which can arise, sometimes because a question is put into context, is the introduction of *differential item functioning*, or *bias*. Test developers do usually want questions to 'function differentially' – that is, to produce different responses from different pupils – since a question which is answered in exactly the same way by everyone taking the test will not serve any obvious educational purpose. But *differential item functioning*, or *bias*, is said to occur when the different responses come from pupils who differ, not in the abilities which the test is attempting to measure, but in some other, extraneous and irrelevant, way.

The most common form of bias is gender based – if a question is placed in a context which is more familiar to girls than to boys, say, then girls may answer it readily while boys cannot understand what is being asked and do relatively badly. So a question relating to netball, for example, might be biased towards girls, while one relating to cricket might favour boys.

In recent years, however, this kind of crude, easily recognisable bias has become less common. If a writer did produce a question which was obviously biased, it would be rejected or amended at

an early stage of the test development process, before any trialling was carried out and statistics were collected. None the less, when statistics relating to questions in a formal trial do become available, they are usually analysed for gender bias. If there are enough of any other clearly identifiable subgroup in the trialling sample for statistics to be significant – if there is a high enough proportion of pupils from a particular part of the country, or with English as an additional language, for example – then data on the performance of this subgroup may also be collected. One or two questions, or parts of a question, may emerge as showing differential item functioning when these studies are carried out, but there is often no obvious reason why these particular questions should have caused a problem. Any question which, on first inspection, looked as though it might be biased will already have been removed, and of two very similar questions, or parts of a question, one may show such differential functioning and the other not. For example, the following question was trialled for a mathematics test:

Put one number in each gap to make the sentences true.

Example

Multiplying by **2** and then by **6** is the same as multiplying by **12**.

- a) Multiplying by **3** and then by **2** is the same as multiplying by \_\_\_\_\_.
- b) Multiplying by **4** and then by **6** is the same as multiplying by \_\_\_\_\_.

The wording, presentation and layout of the two parts of the question are identical. None the less, girls did significantly better than boys (at the one per cent level) in part a), but not in part b). This could have been a random statistical effect, but it still gave the test developers some cause for concern.

When the statistics indicate that some questions in a test are inexplicably biased, it may not be feasible to simply remove them all. A better approach may be to ensure that the test as a whole is unbiased, with as many questions biased towards pupils in one subgroup as in the other.

However, a different situation may arise if a test is designed for a specific, specialised group. For the Key Skill tests in Communication, Application of Number, and Information Technology, for example, test developers were expected to exploit the working environments of the candidates in order to develop questions which really did assess their ability to communicate, apply number, and use IT effectively within the contexts in which they worked. These tests were designed to be appropriate for specific groups of candidates – so in a sense they were 'biased', but the 'bias' was towards candidates who were familiar with the particular context to which the test related. A successful test for trainee gardeners, for example, might have most of its questions set in a gardening context, and one for apprentice hotel workers in the context of a hotel. The test designed for hotel workers might be expected to be 'biased' against the gardeners, and towards the hotel workers, if the question writers had done their job well. In this situation the question writer needs to have a really thorough understanding of the normal working environment of the candidates for whom the test is designed, in order to ensure that the questions are as realistic and meaningful as possible to those particular candidates – although this may make them quite unsuitable for candidates from a different background.



Again, a test developer or a researcher may want to examine the differential item functioning, or bias, of a set of questions, in order to come to a view of the differential performance of the various subgroups. For example, some studies of spatial ability have indicated that boys overall tend to do better than girls on some types of question. If there really are differences in the cognitive style of boys and of girls in some areas of the curriculum, then simply removing all the questions which show any bias will tend to disguise this. As the curriculum backwash takes effect, less emphasis will be placed on those areas in which one group usually does better than another, and curriculum coverage will be skewed. Thus there may be situations in which the test developer or the researcher wants to find biased questions, not in order to take them out of the test, but so that the skills at which pupils in a particular group excel can be identified.

### ***Writing Questions: Key Points***

Four main categories of question are considered: *multiple-choice*, *multiple-response*, *closed-answer* and *open-response*.

Open-response questions are often difficult to mark.

It is difficult to write closed questions which can distinguish effectively between pupils who have *conceptual* understanding, and those who have only *procedural* knowledge.

*Contexts* can make abstract ideas more meaningful.

Test developers must be aware of *differential item functioning*, or *bias*, but the way it is handled may depend upon the purpose of the test.

## **4 Accessibility**

*This chapter focuses more closely on the issue of **accessibility**, which is a key factor in ensuring that a test is a valid assessment of what it purports to assess. The test should be accessible, as it stands and without further modifications, to pupils with the widest possible range of special assessment needs. The resulting improvements to the test are often of benefit in ensuring that it is valid for all pupils, and not only for those for whom the changes have been made.*

*This chapter is likely to be of interest to any reader, whether they wish to select a test for a particular purpose or they are involved in any aspect of test development. The issues raised are central, and should be considered by anyone involved in writing or reviewing test questions.*

### **4a) Should tests be accessible?**

At first glance, it seems obvious that tests should be accessible, if only to ensure that they are valid. If the sentence structure and layout of questions used in a test make it difficult for some pupils to understand what is being asked, then it is not a valid measure of what it purports to measure. As we saw in section 1a), the nature of the assessment itself may prevent some pupils from showing what they know and understand. In that case, the test may be a valid assessment of the pupil's ability to unravel complex sentence structures, or to control their panic in examination conditions – but it is not a valid assessment of the subject being assessed.

However, when the consequences of this argument begin to emerge in tests which are designed to be accessible to the widest possible range of pupils, there may be objections. The situation is not unlike that relating to wheelchair access to public buildings. No one would seriously dispute the necessity for

ramps and lifts in libraries and community centres, at least in theory. But when the ramp is put in place it may be thought to look odd, and to spoil the appearance of the building. Furthermore, able-bodied people may choose to use the lift instead of the stairs inside the building, and so lose the opportunity to take valuable exercise which in the long run would be of benefit to them.

This analogy parallels the conflict which may arise in the context of test development. If the language of the questions is simplified, so that complex, convoluted sentences are replaced by shorter ones with simple structures, then the overall feel of the test may change dramatically. Carefully designed graphics, and perhaps the use of pictures and speech bubbles which 'tell the story' in some questions, may make them much more accessible to a range of pupils, including those who are language impaired or have poor reading skills. Even such minor changes as raising the font size from 12 to 14 or 16 points, and selecting a non-serif font, will have an effect, for poor readers as well as for visually impaired pupils. But some critics, with good eyesight and high levels of literacy and language, may see these changes as undermining the image of the assessment and making it appear facile and patronising. Furthermore, the test developers may be accused of 'dumbing down', and failing to offer linguistically able pupils the opportunity to use and develop the reading and language skills of which they are capable – even if the demand of the questions in relation to the subject being assessed has not been changed at all. Just as the library ramps and lifts may be thought to look odd, and to encourage able-bodied people to be lazy, so tests which are written to be accessible to the widest possible range of pupils may be regarded as patronising, and likely to encourage mainstream pupils to use a lower level of language than that of which they are capable.

Furthermore, most of the things that we can do to make tests more accessible are statistically insignificant, because the

proportion of pupils with the particular disabilities which the design strategies are intended to overcome is too small to have a significant impact on the performance of the group as a whole. Raising the font size, for example, may make all the difference to one or two pupils, but it will neither help nor hinder the great majority. It also puts another constraint on the test developer, who must ensure that the whole question will still fit onto one page, or at the most two facing pages. So it may be argued that a change like this does not make any real difference to most pupils, and is more trouble than it is worth. Critics fear that it could bring the test into disrepute by making it look too easy, so there is pressure to keep the font size small.

But as the *Report of the Disability Rights Task Force* recommends,

Where a policy, practice or procedure places an individual disabled pupil at a substantial disadvantage in comparison with pupils who are not disabled, the provider of school education should be under a statutory duty to make a reasonable adjustment so that it no longer has that effect.

(GB.DfEE, 1999, p.52)

The use of written tests, whether statutory or optional, is a common 'policy, practice or procedure' in our schools. The amendments suggested here may change the overall appearance of the test, making it look easier, but so long as they do not affect the demand of the questions in terms of the subject being assessed they constitute a 'reasonable adjustment' so they should be made. They will make it less likely that disabled pupils will be placed 'at a substantial disadvantage', and so help to ensure that the test is valid for these pupils. In due course there will come a time when even able-bodied people would be surprised and angry to come across a community centre with no wheelchair access, as we become used to the idea that all public buildings should be accessible. In the same way, tests which are

designed to assess what they claim to assess, for the greatest possible number of pupils, will become familiar, and will be accepted as the norm.

#### ***4b) Language and sentence structure***

A number of different factors contribute to the development of accessible tests. The most obvious, perhaps, and usually the first to be considered, is the language, or what Maureen Mobley calls the 'readability' (Mobley, 1987). This goes well beyond issues concerned with vocabulary. As the BATOD (British Association of Teachers of the Deaf) and NATED (National Association for Tertiary Education for Deaf People) booklet, *Language of Examinations*, explains, 'an approach which is 'limited to a consideration of vocabulary is not adequate', because 'sentence structure is at least as important' (BATOD and NATED, undated, p.4).

There are two particular aspects of language which often cause problems for many pupils: conditional phrases, and the use of the passive tense. Both of these make test questions less accessible. As far as possible, questions should be written with a simple subject-verb-object structure, even when this means that there are more words on the page. In the past, a specialised variety of 'examination English' evolved, with instructions and questions which would be found only in written test papers. For example, consider the following two-part question:

- a) *Chalk is added to a test tube containing hydrochloric acid.  
A gas evolves. Identify the gas that evolves.*
- b) *Identify the gas that would evolve if zinc, rather than chalk,  
were added to the hydrochloric acid.*

This question, written in pure 'examinationese', is as much an assessment of language as of science – and as such it would not

be a valid test of science for many pupils. But each part of the question may be rewritten using simpler sentences:

a) *Jane puts some hydrochloric acid and some chalk into a test tube.*

*It makes a gas.*

*What is the gas in Jane's test tube?*

b) *Robert puts some hydrochloric acid and some zinc into another test tube.*

*It makes a gas.*

*What is the gas in Robert's test tube?*

This version has more words than the first, but it is easier to understand. It is composed of simple sentences with very similar structures, so a pupil who has read and understood the situation described in part a) will find part b) quite familiar. All the passive and conditional phrases have been removed.

The simplification of the question is achieved in part through the introduction of two people – and this serves to illustrate one of the main functions of people in a test question. It is sometimes suggested that references to people in questions can make the test seem more 'friendly'. This is also one aspect of the presentation of a question which may be regarded as 'patronising', especially to pupils from a more academic background. Either – or both – of these arguments may be valid – but they should not be relevant to the decision to introduce a named person into a question. Jane and Robert are in the question, adding chalk and zinc to hydrochloric acid, for two reasons. The first is to enable each part of the question to be asked directly, using simple sentence structures. *Chalk is added to a test tube containing hydrochloric acid* is an indirect sentence. *Jane puts some hydrochloric acid and some chalk into a test tube* is direct. The latter is more accessible.

As an alternative to introducing a named third person, questions are sometimes written in the second person – *You put some hydrochloric acid and some chalk into a test tube*. However, the question writer should consider carefully whether this approach will be appropriate to all pupils who might be taking the test. A physically disabled pupil is likely to take part in a school science lesson, but may rely on an assistant to actually put acid and chalk into a test tube. In a test designed for a wide range of pupils it may be more appropriate to avoid the use of 'you' where there could be an implication that 'you' are able-bodied and fully mobile.

The most common reason for introducing named people into a test question is to allow the use of simple, direct sentence structures. A second reason is to signal to the pupils that a new part of the question is starting. In the question shown above, the two test tubes are different and need to be distinguished. Calling them *Jane's test tube* and *Robert's test tube* avoids the need to refer to *the first test tube* and *the second test tube* in a way that is potentially confusing to some pupils.

There is one danger, however, associated with using names in tests. In order to ensure that the test reflects the range of backgrounds of the pupils, a variety of names, both male and female, may be used. But if a name is unfamiliar to some pupils, then it may be taken to be a technical term. This is a particular problem if the name comes at the beginning of a sentence. In the 1992 Key Stage 3 mathematics tests, one part of a question started *Ming collects shapes like these* (KS3 tests, Paper 2, Bands 1 - 4 and 3 - 6). A number of hearing-impaired pupils in a special unit were seen by an observer to reach for their dictionaries, in order to look up the 'technical term' *Ming*.

To minimise this problem, names should not be placed at the beginning of a sentence if this can be avoided. A name in the middle of the sentence will be signalled up to pupils by the

capital letter with which it starts. However, if the main purpose of the name is to avoid the use of a passive sentence – *Jane puts some...* rather than *Chalk is added to...*, then starting the sentence with a name may be unavoidable. *Hydrochloric acid and chalk are put into a test tube by Jane* is no better than *Chalk is added to a test tube containing hydrochloric acid*. In this case, the best option is to use a commonly recognisable name – but inevitably, no name will be equally familiar to all pupils.

While it is essential to keep sentence structures simple in order to develop accessible tests, vocabulary also requires the test developer's attention. The main problem here is not usually with technical language: the pupils' knowledge of technical terms, and their understanding of the concepts to which they relate, will be taught in the classroom and are valid subjects of the assessment. A list of the common technical terms which pupils may be expected to know, and on which they may be assessed, is very useful to the question writer. But more serious problems are likely to arise if colloquial phrases are used in a question. These may have an alternative meaning which could confuse language-impaired pupils or those who are learning English as an additional language. For example, the phrase *carries out an experiment* can conjure up an image of something which is literally picked up and carried out of the room. *Find the result* may seem to be asking the pupil to look for an object.

Non-technical terms which are used to convey the context of a question are particularly susceptible to problems caused by double meanings. A timetable problem in which *Tom caught the bus* led some pupils to imagine Tom chasing after the bus and literally 'catching' it. A better phrase was *Tom got on to the bus* – this is less elegant, perhaps, but it is also less likely to be confusing. Technical terms which are also homonyms can create similar difficulties. For example, one science question could involve a *light object with a low mass*, while another is about a *light source*. In a case like this the two uses of the term *light*



might be unavoidable, but further cues can be given within the question to help pupils to understand which meaning is intended.

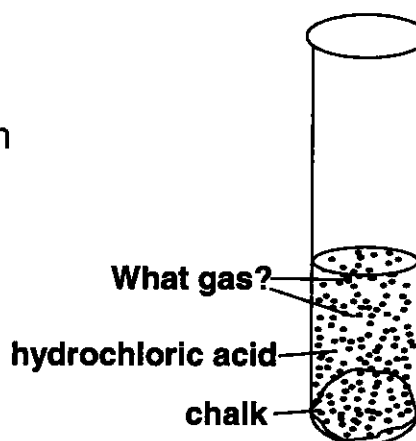
#### ***4c) Test layout***

The language of the test is only one aspect of its presentation, however. The layout, and the way in which information is presented in a variety of forms on the page, must also be considered. A long question may spread over a pair of facing pages – but pupils should not generally have to turn the page in order to complete a question, as they may lose the thread and become confused. Pupils vary in the degree to which they will gather information which is conveyed in different ways: some will read a sentence readily; others will get the same information more effectively from a diagram or chart. For example, the question about the chalk and zinc being put into hydrochloric acid could be presented as:

- a) Jane puts some **hydrochloric acid** and some **chalk** into a test tube.

It makes a **gas**.

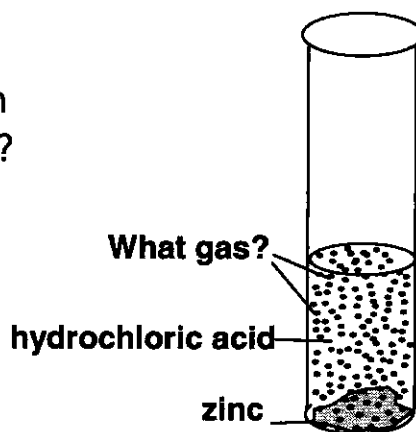
**What** is the **gas** in Jane's test tube?



- b) Robert puts some **hydrochloric acid** and some **zinc** into another test tube.

It makes a **gas**.

**What** is the **gas** in Robert's test tube?



Here, key words from the text are shown in the diagram, and are printed in bold in the question. This will help to ensure that poor readers, and those with limited English, will be able to pick out the essential information in the question by concentrating on the

words which are written in bold. For example, in the sentence just given:

This will help to ensure that **poor readers**, and those with **limited English**, will be able to pick out the **essential information** by concentrating on the words which are written in **bold**.

This approach will support dyslexic readers, who need to focus on the overall picture in order to create a clear image of the situation, and may not be able to process such 'trigger words' as *some, and* or *into* (Davis, 1994, p.23).

If the question is in a context, then pictures, as opposed to diagrams, can help to convey the story and enable pupils to understand what is being asked. The questions about the race given in section 2c), for example, would benefit from a picture of the five racers lined up at the starting line. They should not be shown during the race, however, as any ordering of the runners could serve as a miscue, leading pupils to try to identify the winner, for instance, by looking at the picture rather than using the information given in the table. Furthermore, pictures should be used only if they could actually help some pupils to understand what the question is about: they should never be merely decorative. A picture of Jane looking at the bubbles in the test tube in the question given above, for example, would not convey any information which is essential to the question. Jane is there merely to enable the language to be simplified: she is not directly relevant to the question, and attention should not be drawn to her.

There is sometimes a temptation to try to 'liven up' a test and make it look more attractive by adding pictorial graphics from a ready-made bank after the questions have been developed. This approach is rarely effective. Pictures, like diagrams and people in the tests, are not there to make it more 'friendly', nor to be patronising: they are there to convey meaning which relates to a

specific question. If diagrammatic or pictorial graphics are to be used, then they should be carefully specified within the course of the question development, not added on as an afterthought.

#### ***4d) Modifications and special arrangements***

The aim of any test developer should be to make the test accessible, as it stands, to pupils with the widest possible range of assessment needs. The language and layout of the questions affect their accessibility: the use of simple sentence structures and the avoidance of colloquial phrases help to make a test more accessible to pupils who are not fluent in English or who are language impaired. But we can never develop a test which is perfectly accessible to everyone. Some pupils will have special assessment needs which must be met with special arrangements, modifications to the papers, or both.

Different pupils have different assessment needs, to be met in different ways. Examination boards generally work to a set of guidelines, developed over the years, and other test developers will need to adopt a similar approach. However, the need for modifications to the papers will be minimised by careful development. Furthermore, this will help to ensure that the tests are valid for all pupils, not just for those with special assessment needs.

For example, a good way to ensure that the language of a test is as clear as possible is to have the whole test Signed, as for a deaf pupil, and then to translate the Sign as directly as possible back into standard English. Complex sentence structures in English will be lost when the test is translated from Sign. For instance, consider these mental arithmetic questions:

*Subtract sixty-eight from ninety-three.*

*Take away sixty-eight from ninety-three.*

*Ninety-three subtract sixty-eight.*  
*From ninety-three, take sixty-eight.*

The Sign for each of these could be translated as *Ninety-three subtract/take sixty-eight*. Knowledge of the technical term *subtract* might be assessed in the spoken English test, but the other variations represent differences in sentence structure rather than vocabulary. The spoken English and the Sign versions of the test should be as close as possible, so the spoken English test should use *Ninety-three subtract sixty-eight* or *Ninety-three take (away) sixty-eight*, rather than one of the less direct sentence structures. But using this form of the question will help to ensure that for hearing pupils, as well as for deaf, the assessment measures the ability to subtract sixty-eight from ninety-three, rather than the ability to unravel an indirect sentence.

In general, then, the well-established dictum *If it's good for special then it's good for mainstream* applies to tests as to any other aspect of the curriculum. The logical inverse, *If it's bad for mainstream then it's bad for special*, is also true – a badly structured question will be difficult for many mainstream pupils to access, but it is likely to be quite impossible for some pupils with special assessment needs, for whom it will be completely invalid.

But there is one type of modification to which this does not apply. Modified papers for blind pupils, whether Brailled or with modified large print, are likely to hinder rather than help many mainstream pupils. These papers are largely stripped of the diagrams, graphics, boldening and other layout features which may serve as prompts for sighted pupils, but cannot be accessed by visually impaired pupils. Again, a visually impaired pupil cannot 'scan' a table of information or a diagram: the information must be taken in piece by piece, and held in the memory until the whole picture is built up. This is a difficult and time-consuming process, so, as far as possible, key information

should be presented entirely in words. Any diagram or table which cannot be replaced by a succinct description should be simplified as much as possible, so that a pupil who is 'reading' it a bit at a time, by touch or by eye, can grasp the main features more easily. This being the case, whereas modifications made for most pupils with special educational needs are likely to be appropriate for mainstream pupils, and may even be of some benefit, those made for visually impaired pupils are often unsuitable for more general use.

### ***Accessibility: Key Points***

The aim of the test developer should be to make the test *accessible*, as it stands, to pupils with the widest possible range of special assessment needs.

Wherever possible, a simple subject-verb-object sentence structure should be used.

A named person may be described carrying out an action in a question. This may enable the test developer to use simple, direct sentence structures.

The separate parts of a question may also be signalled up with references to different named people.

Key information given in bold, and the use of pictures and labelled diagrams, can help to make questions more accessible.

Modifications made for hearing- or language-impaired pupils should contribute to the development of the mainstream test.

Modifications made for visually impaired pupils are not likely to be suitable for mainstream pupils.

## 5 The Test Development Cycle

*The overall cycle of test development, from the first specification to the marked test, is a complex process. Questions must be drafted, tried out with a few pupils, and then, usually, redrafted. Experts must scrutinise the test, and comment on such issues as its accessibility, subject accuracy and validity. Mark schemes must be developed and trialled, and these too are likely to be amended several times. Each part of the test development interacts with the others to create a dynamic process.*

*In this chapter, some aspects of the organisation of question writing are considered. The difference between **formal** and **informal trialling** is discussed, and their key functions are explained. The development of robust, detailed mark schemes, which offer sufficient guidance for the marking of a wide range of possible types of answer to open-response questions, is also discussed.*

*This chapter is likely to be of interest to people involved in any aspect of the development of formal tests or examinations, including question writers or reviewers and those who oversee the test development process as a whole.*

### **5a) Organising question writing**

If the task of drafting the questions is to be shared among a number of writers, then this must be organised carefully. Each writer may be asked to produce a certain number of ‘marks’ worth of questions. Different writers should be asked to address different parts of the test specification, in order to ensure a balanced coverage. However, to allow writers to exploit a rich question context as thoroughly as possible, it is best to allocate to each a proportion of marks which may cover any aspects of the specification.

Experts with experience of teaching pupils with a variety of special educational and assessment needs should be involved early on in the process. Teachers of hearing-impaired pupils are of particular importance here, as the advice they offer on the use of clear, plain language should feed into the development of the mainstream questions. This will help to ensure that the papers are accessible as they stand to the widest possible range of pupils. However, the questions are bound to change significantly as the test development cycle goes on, so the advisers should be asked to comment on the drafts throughout this process.

If the test is being developed in two languages, then parallel development is likely to be far more effective than translation. Fluent speakers of all the languages in which the test is to be developed should be involved in the test development process from the beginning. This will help to ensure that both the contexts of the questions and the way that they are expressed are equally appropriate in both languages. If the advice and guidance of speakers of each language is not sought early on, questions may be developed which work perfectly well in one language and culture but are meaningless or inappropriate in another. To take a simple example, *How many sides does a pentagon have?* is a reasonable, if uninspired, question in English. In Greek or Turkish, however, it translates as *How many sides does a five-sided shape have?* Clearly, this question should be rejected as being ineffective if it may be asked in one of these languages. Again, a question developed for a science test asked *Some animals hibernate. What does this mean?* Since the test was intended for use in Irish schools, and the term commonly used for *hibernate* in Irish-medium primary schools translates directly as *sleeps in winter*, the question could not be used. Similarly, a question about a school timetable which did not mention Welsh language lessons was unacceptable in a test designed for pupils in England and in Wales, since Welsh is a



compulsory subject in all schools in Wales. To develop and trial such questions only to drop them later when they prove untranslatable or inappropriate is not an efficient use of time and effort.

As questions are developed they must be trialled. There are essentially two types of trialling: informal, when the immediate response of a small number of pupils to a question is observed; and formal, when detailed statistical data are collected. These two aspects of trialling are discussed separately below.

### ***5b) Informal trials***

In an *informal trial* a small number of pupils try out a draft question very informally, perhaps discussing their work with each other, or with their teacher or the test developer. Informal trialling of individual questions sometimes gets squeezed out in the pressure to get a draft test ready in time for the formal trials. But over the years test developers have often found that showing a draft question to half a dozen pupils in one classroom pays quite disproportionate dividends. If this crucial stage is missed out, and a question goes forward to a formal trial without informal trialling, it may have some 'obvious', but unnoticed, flaw which will appear from the results of the formal trials. This is an expensive waste, since formal trialling costs far more than informal.

For example, a question in a primary school science test described how some children investigated the conditions under which bean plants grew. One part of the question focused on the need for light. The question writer used 'Nick' and 'Sarah' to convey the idea that two different sets of conditions were set up:

Nick puts his seedling on the **window sill**.

Sarah puts her seedling inside a **dark cupboard**.

Whose seedling will **grow better**?

**Explain your answer.**

The draft mark scheme required pupils to identify Nick, and to make some reference to *light* in their explanations. However, during an informal trial of this question, a problem emerged in connection with the science itself. While it is true that plants need light to grow healthily, in the early stages a seed in a dark place may germinate and then grow very fast, becoming long and spindly in its search for light. In the trial one pupil wrote *Sarah's seedling would grow taller in the dark cupboard because it would be looking for the light*, taking *grow better* to mean *grow taller* rather than *grow more healthily*. Another pupil wrote *Nick's seedling will grow better because it is on the window sill*. This response indicates quite clearly whose seedling will grow better, and it offers an explanation for that growth – so it was a reasonable answer to the question as it had been set. However, it did not refer specifically to the need for light, so it did not meet the requirement of the draft mark scheme.

In order to overcome these problems, the question was rewritten as:

Nick keeps his plant on the **window sill** for four weeks.

Sarah keeps her plant inside a **dark cupboard** for four weeks.

In **which place** will the plant will be **healthier**?

**Why** will it be healthier?

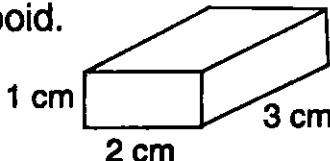
The reference to *seedlings* was changed to *plants*, and the question stated that the plants were kept in position for four weeks – too long for a bean seedling to maintain its first spindly growth in the dark. This helped to avoid misreading pupils into giving responses like that of the first pupil above. The question also became more focused: it asked why the plant in one position would be healthier, rather than asking for a general explanation which could provoke a response referring back to the position of the plant, like that of the second pupil in the informal trial.

Informal trialling serves another important purpose when the questions being developed are multiple-choice. A good multiple-choice question relies for its effectiveness on having at least three good *distracters* – plausible, but definitely incorrect, answers which can be offered as options in the question. These often arise from common misconceptions, or types of error which pupils are likely to make. A good way to identify these errors in the context of a particular question is to set it as an open-response question to just a few pupils, and then to work out, perhaps by talking it through with the pupils afterwards, where different wrong answers came from. Frequently occurring misconceptions can then be incorporated into the distracters in the multiple-choice version of the question.

Informally trialling questions as open-response, rather than multiple-choice, and discussing the answers afterwards with the pupils, may also help to avoid questions which pupils may answer correctly for the wrong reasons. These are a test developer's nightmare, and can never be totally ruled out. For example, the following question was informally trialled as an open-response question:

Find the **volume** of this cuboid.

Answer: \_\_\_\_\_ cm<sup>3</sup>



A pupil wrote:

$$1 + 2 + 3 = 6$$

This, of course, gave the correct answer – but as the result of an erroneous calculation. If the question had been trialled only in its multiple-choice version, then this serious weakness might not have been recognised.

A similarly flawed question relating to the data on the race which was given in section 2c) would be:

**Which runner came third?**

There were five runners in the race, so the third last was also the third in the race. This being so, a candidate who made the common error of associating the highest number on the table with the winner would get the correct answer.

Finally, effective informal trialling reduces the need to trial a great many more questions than will actually be required for the final version of the test, so it is likely to be much more economical in the long run. If more than double the number of questions needed for the final version are being formally trialled, then it is worth considering the situation carefully. It may be better to concentrate on developing a smaller number of really good questions, and to amend them where necessary, rather than formally trialling a lot of material which there has not been time to trial informally.

***5c) Formal trials***

Formal trials are those from which meaningful statistics may be gathered. They fall into two types: *individual question trials* and *whole-test trials*. In an individual question trial, statistical information is gathered on the performance of pupils in relation to each mark in each question. This information helps the test

developers to decide which questions, or which parts of a question, should be included in the final version of the test.

In an individual question trial, the questions are collected into test papers, to allow them to be administered to large samples of pupils, but these test papers are likely to be unbalanced. They may have more questions focusing on areas of the curriculum that are hard to assess, to allow for the inevitable wastage of questions in these areas. Such a trial therefore normally includes more questions than will eventually be needed, and also some which are carrying more marks than can actually be used. For example, two similar questions focusing on the life cycles of the frog and the newt might be developed. Statistics could then be collected on pupils' performance on both these questions, and pupils' work could be scrutinised. The two topics would be likely to fall into the same area on the coverage chart, however, so only one of the questions would be carried forward to the final test.

An individual question trial also gives developers the opportunity to trial slightly different versions of the same question, and to collect data on the differences that may be caused by small variations in the wording or layout of a question. For example, in the course of development of the 1998 Key Stage 3 mathematics tests, two versions of an algebra question were trialled to see the effect on the pupils' performance of different types of graphic when the wording of the question was kept the same.

### ***Marbles: Introduction***



Jane and Tom each have a bag of marbles.

Each bag has the **same** number of marbles inside.

You cannot see how many marbles are inside each bag.

Call the number of marbles inside each bag  $x$ .

### **Question with abstract diagrams**

- (a) Jane puts **5 more** marbles **into** her bag.



Write an expression for the  
total number of marbles in Jane's bag now.

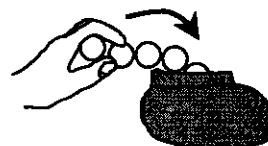
- (b) Tom takes **2** marbles **out** of his bag.



Write an expression for the  
total number of marbles in Tom's bag now.

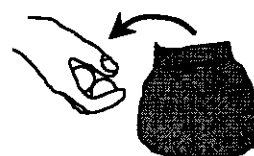
**Question with pictures of a hand**

- (a) Jane puts **5 more** marbles **into** her bag.



Write an expression for the total number of marbles in Jane's bag now.

- (b) Tom takes **2** marbles **out** of his bag.



Write an expression for the total number of marbles in Tom's bag now.

It was found that showing the picture of a hand moving the marbles in and out of the bag improved the performance of some groups of pupils, so this was the version that was taken forward to the final tests. (See Clausen-May, 1998 for a more detailed account of the development of this question.)

However, a balance must be maintained between the number of parallel versions of the same question, different questions assessing the same topic, and different questions assessing different topics. This will ensure that, at the end of the day, there is enough choice available to enable the test developer to put together a complete, coherent, balanced final test.

While an individual question trial offers opportunities for experiment, and is intended to support the process of question development, the whole-test trial is much more restricted. Statistics relating to the individual marks in each question will be collected, but the focus of the trial is on obtaining statistics relating to the pupils' performance on the test as a whole. This

being the case, the test papers should be as close as possible to the final versions. If the purpose of the trial is to set population norms, for example in a standardisation trial (see section 6c), then there should be no further amendments to the questions once the statistics have been collected.

If *cut scores*, the number of marks needed by a pupil for the award of a particular grade or level (see section 6e), are to be based on the statistical results of a trial of this sort, then the trialled papers should again be regarded as final. Developers sometimes try to guess whether a change made to a question after the whole-test trial will make it easier or harder, so that this can be allowed for in the statistical analysis, but this process is always risky. Even rearranging the order of the questions in the test, let alone making changes to their wording, layout or presentation, may have a significant effect on the statistics relating to the pupils' performance.

#### ***5d) Developing robust mark schemes***

Formal trials have at least two distinct purposes. The first has already been discussed: statistical information is collected about each individual mark, and also, in the whole test trial, about the complete test. But this is only one of the outcomes of formal trialling. The other is the development of robust mark schemes.

Different types of question need very different types of mark scheme. A good multiple-choice question depends upon having plausible, but definitely incorrect, distracters. The mark scheme itself is very straightforward. For closed-answer questions the mark scheme may also be relatively easy to write – although there are pitfalls even here. But the real challenge comes with writing mark schemes for open-response questions. Early versions of these are bound to be skeletal and imprecise, with any number of possible alternative responses left uncovered.



The final mark schemes cannot suffer from such weaknesses: they must be based on a careful scrutiny of real responses given by real pupils under realistic test conditions.

It is often harder to write good mark schemes for open-response questions than it is to write the questions themselves. The two must go hand in hand: the art is in writing *markable* open-response questions. Such questions are always likely to throw up unforeseen answers which need to be covered. If the mark scheme is not published in advance – or not published at all – then a last-minute decision may be made at a markers' meeting to deal with a response which does not match that given in the mark scheme, but which may, none the less, be worthy of credit. But with the increasing stress on accountability, unpublished mark schemes are becoming less acceptable. Following all national tests, for example, scripts are returned to schools along with the mark schemes which have been finalised before the tests were taken. This puts much greater pressure on the test developers, and may militate against the use of open-response questions.

A robust mark scheme must cover all possible responses to the question, indicating clearly which are acceptable and which are not. It should include a generalised description of the correct response, followed by one or more examples – not an exhaustive list, but an indication of the sorts of responses pupils may give. But at the same time, there is always pressure to keep the mark scheme simple. For example, a mark scheme such as:

<b>Correct Response:</b>	Indicates the correct length of time
<b>Example:</b>	18 seconds

may draw objections. *If the answer to the question is '18 seconds', then give that as the Correct Response and have done.* Markers may not see the necessity for a wordy generalised description of the correct response – *Indicates the correct length of time* – with *18 seconds* given as an example. Surely *18*

*seconds* is not just an example – it is the correct answer, and that is that!

But is it, always? What about *eighteen seconds*? Or *0.3* (or *nought point three* or *zero point three*) *minutes*? Or  $\frac{3}{10}$  (or *three-tenths*) *of a minute*? Depending on the nature of the question, and the different methods pupils may use to answer it, any of these responses might turn up – and at least some of them might be worth a mark. Catch-all phrases such as *Accept equivalent answers* may serve well enough for most responses to a straightforward question, but there is always a danger that markers will come to different decisions about what is, and what is not, 'equivalent'. Giving one, precise, definitive answer to a question, even to a closed question which to all appearances has got a definitive answer, is unwise. The chances are that someone, somewhere, will find another answer which could be just as good.

With an open-response question, however, the problems are much greater. Here, a general statement summarising the essence of a correct response is essential. This should be followed by examples of creditable answers, taken from real pupils' work culled from the trials. These examples are not 'model answers' – they are not detailed, correctly expressed responses to which any marker would award full marks with no hesitation. Rather, they should show the most common types of response for which marks may be awarded, expressed in the sort of language and with the kind of layout which pupils actually use. They should be distinguished from answers which just miss being awarded the mark, and the borderline should be established.

For example, a mark scheme was developed for the question about growing plants which was given in section 5b):

Nick keeps his plant on the **window sill** for four weeks.

Sarah keeps her plant inside a **dark cupboard** for four weeks.

In **which place** will the plant will be **healthier**?  
**Why** will it be healthier?

The mark scheme gave a general description of the required response, followed by some examples, ranging from the fairly comprehensive *On the sill, because the plant will be in the light* to the minimal *Window gets sun*:

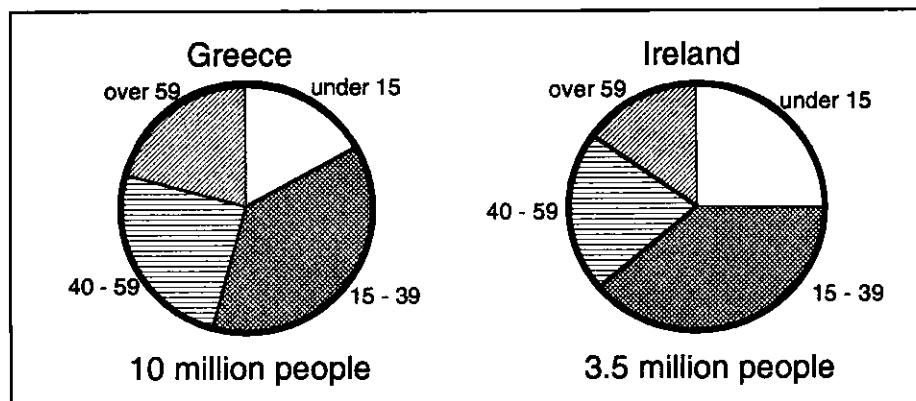
Correct response	Notes
<p>Indicates the plant on the window sill, and explains that plants need light to grow, e.g.  <i>On the sill, because the plant will be in the light.</i>  <i>Nick's plant will get more sunshine.</i>  <i>Window gets sun.</i></p>	<p>The position of the healthier plant need not be identified, so long as an appropriate explanation is given,  e.g.  <i>Sarah's plant won't get enough light.</i>  <i>Plants want light.</i>  <i>Beans won't grow in the dark.</i></p> <p>Do not accept an inappropriate explanation if a plant has been identified,  e.g. do not accept  <i>Nick put his plant in a dark place, so it won't grow so well.</i>  <i>Sarah's plant will grow better, because it will get more light.</i></p>

The *Notes* in the right-hand column of the mark scheme cover a number of further variations. They indicate that a response may be accepted even if it does not directly answer the first part of the question *In which position will the plant be healthier?* so long as it shows a good level of understanding of the part light plays in plant growth, which is the real focus of the question. The second example given in this column, *Plants want light*, also shows that the test developers have decided that, in this case, the term *want*, rather than the more technically correct *need*, should be accepted.

On the other hand, the decision was taken to reject responses such as *Nick put his plant in a dark place, so it won't grow so well*, because the explanation is inappropriate to the selected plant. It could be argued that this answer demonstrates a clear grasp of the scientific principles involved, and the reference to Nick is a mere slip, so the response (and others like it) should be accepted. Such an approach is perfectly feasible, and it could well be taken. The important point, however, is that a decision one way or the other must be made, and must be specified clearly in the mark scheme. It is up to the test developers, in consultation with the relevant advisers, to decide exactly where the borderline lies between responses which are, and those which are not, worthy of credit. Such decisions should not be fudged when the mark scheme is being developed.

The distinction between minimally acceptable and minimally unacceptable answers to an open-response question is crucial – and it may be hard to explain to the marker. For example, in the report on the 1998 Key Stage 3 mathematics tests, a question, *Ages*, which involved the analysis of some demographic data relating to two countries, Greece and Ireland, was discussed (QCA, 1998, p.23). The information in the question was given on two pie charts: these indicated that the populations of Greece and Ireland were 10 million and 3.5 million respectively. Each pie chart was divided into sectors representing the proportion in

different age bands in the two countries. The sector labelled *under 15* took up about a quarter of the Irish chart, but only about a sixth of the Greek chart. Thus the Irish *under 15* sector was bigger than the Greek.



In part (c) of the question, pupils were asked to explain why the statement *The charts show that there are more people under 15 in Ireland than in Greece* was wrong. The mark scheme required pupils to indicate that *the total number of people in each country needs to be taken into account*. Several examples of correct responses were given, including *Ireland has **only** 3.5 million people*. But the mark scheme also gave the very similar response *Ireland has a population of 3.5 million* as an example of an *incorrect* answer, since this merely restates information given in the question without bringing out the comparison with the total population of Greece. The essential difference between these two responses is in the word *only*: this implies a comparison, and so it gets the mark. It is by no means a good response, but it is – minimally – acceptable, and worthy of credit. Distinctions like this, between marginally acceptable and marginally unacceptable responses, must be made clear in the mark scheme, and must be based upon a careful scrutiny of pupils' work from the formal and informal trials.

We have seen that one of the purposes of informal trialling is to try to spot questions which can be answered correctly for the

wrong reasons (see section 5b). When closed-answer or open-response questions in the formal trials are marked, pupils may again show by their working that their reasoning was incorrect, even though their answer is correct. When this happens, the question must be amended, or scrapped. None the less, a few such questions are bound to slip through the net occasionally. When a mathematics test is being developed, this can lead to an attempt to outlaw any responses which are not supported by evidence of a correct method.

But this is not always as simple as it seems. The pupil's working to find the volume of the cuboid given in section 5b) was clearly incorrect, but another pupil showed the same response, but with symbols which could have been addition or multiplication signs. The marker could not decide which they were – and in these circumstances it would be difficult to justify withholding the mark on the grounds that the working could be incorrect, when the answer is correct.

Requiring pupils to show their working in a mathematics test may also tend to encourage inappropriate written methods when a mental strategy would be more efficient. For example, asked to find the cost of six pens at £1.99 each, many pupils would use the efficient mental strategy of multiplying £2 by six and then subtracting 6 pence. If they were required to write down their working, however, then some pupils would be likely to set out their calculation as:

$$\begin{array}{r} \text{£}1.99 \\ \times 6 \\ \hline \end{array}$$

They might go on to use a standard written algorithm, relying on rote learning rather than an understanding of the computation involved.

Furthermore, the requirement that a correct method be shown places a heavy burden on markers who are expected to check the working of all pupils, even if they have got the answer to the question correct. There is always a danger that a perfectly good, but unfamiliar, method, will be used by a pupil, perhaps a recent arrival who has been taught to lay out their work using different conventions. In the end, the difficulty of marking 'methods' may lead to an increase in the number of multiple-choice or multiple-response questions. These are still prone to incorrect reasoning, but are better able to mask the problem since the marker will not be attending to the pupils' working, and indeed may not even see it. As Paul Black observes,

Some studies have shown that up to a third of pupils who choose a correct response may do so for a wrong reason.

(Black, 1998, p.83)

In a multiple-choice test, however, the marker has no way of telling when this has occurred, so the problem is disguised, although it may be reflected in the statistics when the question is trialled.

### ***5e) Cycles and spirals***

The description of the test development cycle given here is somewhat linear. However, the way it is carried out is rather more complex, with cycles within cycles feeding the results of each trial and consultation back into the development process. If informal trialling shows that pupils are misled by some aspect of the wording or presentation of a question, then this must be revised. If the question still fails to provoke a meaningful (even if incomplete or incorrect) response, then the aim of the question itself may need to be reconsidered. When the question writer is unclear about what it is that is being assessed, the pupils are also

likely to go astray. Questions do not necessarily 'work' or 'not work': there are many shades of grey in between!

The advice of teachers who have worked with pupils with a range of special educational needs will help to ensure that the questions are accessible, but this too is likely to take several iterations before the best possible wording and layout are found. Questions should have undergone sufficient informal trialling before large-scale, formal trials are carried out to ensure that few questions have to be rejected outright – but if a question is simply unmarkable, for example, then it will have to be radically rewritten or dropped. Any rewriting requires further informal trialling, and another check with the language consultants to ensure that nothing has been done to reduce accessibility. Thus the process cycles round, with back-currents and eddies rather than a simple, linear flow.

### ***The Test Development Cycle: Key Points***

Small-scale, informal trials are quick, cheap, and invaluable.

Large-scale, formal trials are required to obtain meaningful statistics. *Individual question trials* give statistics relating to each mark in each question. *Whole-test trials* give statistics relating to individual questions, but also to the test as a whole.

The development of robust mark schemes also relies on extensive trialling, both formal and informal.

Mark schemes for open questions must distinguish between responses which just do, and those which just do not, get the mark.

The test development cycle is a spiral rather than a straight line, with each trial and consultation feeding back into the process of test development.



## 6 Statistics for Test Users

*It would not be feasible, in a book like this, to cover all the highly specialised statistical knowledge associated with test development. This must be left to the expert statisticians. None the less, while those statisticians can work out the precise values of the 'reliability coefficients' or the 'confidence bands', writers and users of the tests need to understand what such statistics actually mean.*

*In this chapter, the key concepts of **validity** and **reliability** are examined again, this time from a more statistical point of view. The use of **standardised scores** is then discussed, and the idea of a **confidence band** is explained. Finally, **cut scores** are examined and some of their effects are considered. All of these statistical ideas are likely to be of general interest, to the users of formal tests or examinations as well as to their developers and reviewers.*

### **6a) Another look at validity**

We have seen how the *validity* of a test – the extent to which it measures what it purports to measure – is partly dependent upon its accessibility. This is one aspect of what is called the *content* validity of the test. However, a more statistical approach gives us two different ways to view validity: *predictive* validity and *concurrent* validity.

As Ian Schagen explains in his article on the statistics of tests, 'Testing, testing, testing',

'Content validity' means that the content of the test addresses the area of interest to the assessment. 'Predictive validity' means that the test gives valid predictions of another relevant outcome which may be judged statistically. If, for example, a test is supposed

to predict GCSE performance, we could collect data and decide how well it did this using statistical tests.

(Schagen, 1999, pp.28 - 9)

So we could say, on the basis of such 'statistical tests', that *There is such and such a probability that a pupil who scored so and so on the test will achieve whatever at GCSE*. Such a predictive test is a useful tool – although it has its dangers, as it may lead to self-fulfilling prophecies, with pupils being allocated to high or low sets with different expectations and curricula.

Predictive tests may also create a backwash effect on the curriculum. To take an absurd example: supposing it was found that, for whatever reason, children who had pointed ears got better test and examination results than those with round ears. If this were taken seriously, and decisions were made on the basis of the 'predictive validity' of this 'test', then it would not be long before special devices were sold in the chemist's shop, or made available through the school supplies catalogue, to help to ensure that children's ears grew in a more pointed shape. If performance on ability A is used to predict performance on the highly valued ability B, then eventually teachers (and perhaps parents) will teach ability A. Thus a test which was intended to be entirely 'summative' – to provide evidence of a pupil's current level of performance – would become 'formative', and would be used to guide the teacher's actions – in this case quite inappropriately. So, for example, 'practice' non-verbal reasoning tests can be found in bookshops and in school publishers' catalogues. Non-verbal reasoning tests may be good predictors of various forms of academic success – but making children better at doing the tests will not make them better at doing any of the things which the tests are used to predict. None the less, parents and teachers do buy and use the 'practice' tests, so non-verbal reasoning tests are, to some extent at least, having an impact on the curriculum. In this sense they are 'formative' and not just 'summative'.

Like predictive validity, concurrent validity also may be measured statistically. As Ian Schagen explains,

One way of determining the validity of a new test of ability X, if we have an existing test of X, would be to compare results between the new and the old tests.

(Schagen, 1999, pp.28)

If they correlate well – if pupils who do well with one test do well with the other, and those who score badly do so on both – then we may argue that the two tests are measuring the same thing. But clearly, the concurrent validity of the new test will be only as good as the test we are comparing it with. A lot of care may have been taken to ensure that the new test has a high level of content validity, with accessible language and good layout and presentation. The established test may have a lower level of content validity, if only because it is out of date, with inappropriate question contexts and dated, formal language. The level of correlation between the pupils' performance on the two tests – the concurrent validity – is likely to reflect their content validity.

So concurrent validity does not guarantee content validity. As was discussed in section 1b), a closed test which assesses procedural knowledge is much easier and cheaper to develop and mark than a test which assesses conceptual understanding. Pupils' performance on the former may correlate well with the latter, but that does not make it a test of conceptual understanding. And as the curriculum backwash takes effect, the use of a test which does not assess the pupils' understanding may tend to undermine teaching and learning in the classroom.

Both predictive and concurrent validity are, at least in part, statistically defined. Another, broader approach involves what is called *construct validity* (Messick, 1989). Messick makes the distinction between *content validity* and *construct validity*,

suggesting that content validity is about behaviour, while construct validity is about cognition. As he explains,

Content validity is evaluated by showing how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn.

(Messick, 1989, p.16)

Any test is composed of just one selection from the set of all the possible questions which could be asked about the subject being assessed. The validity of the test depends in part on how well the selection represents the whole set – on whether it covers all aspects of the subject, or is heavily loaded with questions addressing one part of the curriculum rather than another. So,

Content validity is based on professional judgements about the relevance of the test content of a particular behavioural domain of interest and about the representativeness with which item or task content covers that domain.

(Messick, 1989, p.17)

On the other hand,

Construct validity is evaluated by investigating what qualities a test measures, that is, by determining the degree to which certain explanatory concepts or constructs account for performance on the test.

(Messick, 1989, p.16)

So construct validity may encompass other forms of validity, and is at least in part a philosophical concept. As Messick explains,

There is often no sharp distinction between test content and test construct... content-related inferences and construct-related inferences are inseparable.

(Messick, 1989, p.36)

But in any case, according to Messick,

What is to be validated is not the test or observation device as such but the inferences derived from test scores or other indicators – inferences about score meaning or interpretation and about the implications for action that the interpretation entails.

(Messick, 1989, p.13)

So, in a sense, all tests are seen as primarily formative, and their validity depends upon the consequences – including any curriculum backwash effects which they may create – which follow from their use. Since, as we saw in section 1b), *What You Test Is What You Teach*, test developers have a responsibility to ensure that what is taught as a result of what is tested is educationally sound, and supports good practice in teaching and learning.

### ***6b) Reliability as a statistical concept***

While validity is, at least in part, a philosophical concept, the *reliability* of a test can be statistically defined in a variety of ways. The reliability is a measure of the extent to which the results of a test are reproducible. Any one test consists of one particular set of questions – but a different set, chosen out of the universal set of all possible questions covering the same topics, would probably give different results. As Paul Black explains,

Any practicable examination can sample only a limited number of the possibilities, and it is then important to be able to estimate how inaccurate the result might be because of this limitation. One way of doing this is to analyse the internal consistency of pupils' responses. If pupils each respond with about the same level of success to different questions, and if these questions are a fair sample of the range of

possible questions, then one can be confident that their overall score reflects accurately what these pupils might have attained over the full range of possibilities. A simple way to explore this is to divide any test paper into two halves and check the agreement between the marks for the separate halves. More complex techniques are equivalent to doing this, but they go further by averaging the result of making the split into two halves in every possible way.

(Black, 1998, p.40)

Using a method like this, the statistician can then calculate the *reliability coefficient*, a number between nought and one which is a measure of the test's reliability. The higher the reliability coefficient the more reliable – in the statistical sense – the test will be.

Another approach to measuring the reliability of a test is through what is called a test-retest correlation. The same pupils are given the same test twice, and their scores are compared to see whether their performance is similar on both occasions. To be useful, the results of a test need to be reasonably consistent over time.

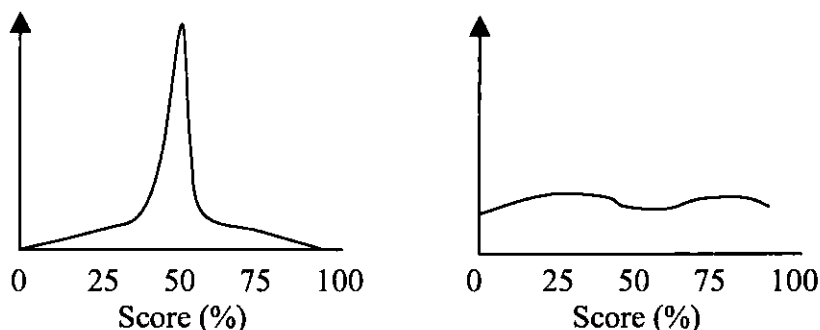
However, it must be remembered that the fact that a test is highly reliable tells us nothing about its validity. This is nicely demonstrated by an apocryphal story relating to a test composed of 32 questions which was given twice to the same group of pupils. The test-retest correlation was perfect: every pupil got exactly the same mark the second time as they had got the first. The test developers were startled: this was most unusual, but their test had apparently been proved to be utterly reliable. Unfortunately (or perhaps it was just as well really) someone spotted that there had been an error when the data from the tests were entered into the statistical analysis program. The column which should have contained the pupils' scores instead contained

the day in the month when they had been born. This did not change when the pupils took the test for the second time, so this 'test' – of the day on which pupils were born – was, indeed, reliable. But it was not a valid test of anything which the test developers might have wanted to assess.

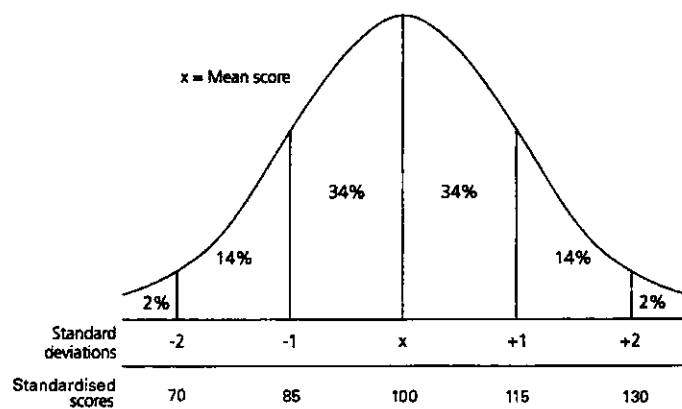
### ***6c) Standardised scores***

When a test script is marked, the number of marks awarded to the pupil is found. But this figure – the raw score – tells us little on its own. Tests carry different numbers of marks. One test may be marked out of 30, and another out of 50. So a score of, say, 25 would mean quite different things on the two tests.

But even if two tests are marked out of the same total, one may be much harder than the other – so a mark of 25 out of 30, for example, might indicate a performance which is well above average on a test with 30 difficult questions, but about average on a test with 30 easy ones. Again, two tests might have the same means, but quite different distributions – so, for example, nearly all the pupils might get 50 per cent on one test, while they were distributed fairly evenly on the other. In that case, a pupil getting 40 per cent on the each test would have done worse than nearly everybody else on the first, but as well as nearly half of the other pupils on the second one. Distributions like these are shown in the graphs below.



This being the case, the *raw scores* (the number of marks which pupils get) on a test may be converted to *standardised scores*. These are based on the performance of a large, representative sample of pupils in a *standardisation trial*. These pupils take the test, and the mean and standard deviation of their raw scores is found. This set of scores is then converted to standardised scores, which are usually set in such a way that the mean standardised score is 100, with a standard deviation of 15. In other words, an average pupil in the standardisation trial is given a standardised score of 100, and two-thirds of the pupils are fitted in between standardised scores ranging from 85 to 115 (or  $100 \pm 15$ ).



So the majority (two thirds) of the pupils in the standardisation trial will score between 85 and 115. A pupil who gets a standardised score of over 100 is performing above average – and one who gets a score over 115 is doing very well indeed. Converting the raw scores to standardised scores allows teachers to compare pupils' performance on different tests which may have different levels of difficulty, carry different numbers of marks, or have different score distributions.

Furthermore, the test may be standardised in such a way as to allow the ages of the pupils involved in the trial to be taken into



account. As the Teacher's Guide to one standardised test, *Mental Mathematics Test 11*, explains,

An older pupil may gain a higher raw score than a younger pupil, but have a lower standardised score. This is because the older pupil is being compared with other older pupils in the reference group, and has a lower performance relative to his or her own age group.

(Clausen-May *et al.*, 1999, p.23)

So a *standardisation trial* is a statistical exercise, designed to produce a set of data. Using this data, the statistician can draw up a table which will enable the teacher to convert any pupil's raw score into a standardised score, which will indicate how the pupil has performed in comparison with other pupils in the standardisation sample.

The statistical exercise of a standardisation trial is sometimes confused with standardised administration, which may specifically prohibit the use of special arrangements even for pupils with special educational and assessment needs. However, as we have seen, this is likely to make the test results invalid for some pupils. Rather, pupils with special educational needs who take part in a standardisation trial should be given the same level of support as they normally receive in the classroom, and would be expected to receive during any formal assessment. If this is not practicable, then it may be better if such pupils are not included in the standardisation trial. However, the number of pupils needing special support who will be included in a representative sample is so small that their results are unlikely to have much effect on the overall figures collected from the standardisation trial.

#### ***6d) Measurement error, true scores, and confidence bands***

Statisticians sometimes refer to what they call *measurement error*. As Ian Schagen explains,

Random variations in test results, unrelated to individuals' abilities or other factors we are trying to measure, are known as measurement error.

(Schagen, 1999, p.28)

But as the reader will have gathered from the sections on question writing and test development, creating questions which measure only 'individuals' abilities or other factors we are trying to measure' is not easy. Aspects of the test itself, the way in which the pupils respond, the test conditions, and so forth may introduce measurement error. This measurement error will affect what is known as the *true score*. The true score is not the actual outcome of a real test: rather, it is what the pupil 'should' have scored if all the 'random variations' resulting from the measurement error could have been avoided.

The concept of a true score depends upon the idea of measurement error, and this can help to explain what is meant by a *confidence band*. Converting raw scores to standardised scores allows us to say that Sally got a standardised score of 103 on this test, and Jenny got a standardised score of 97 on that one, so Sally has performed a little above, and Jenny a little below, the average for the standardisation samples. But with such a small difference, can we say that Sally is better than Jenny? The confidence band allows us to get a feel for the relationship between a pair of standardised scores like these. It takes account of measurement error, and as Ian Schagen indicates, 'it allows us to quantify the real underlying uncertainty in any test result'.

The *Teacher's Guide to the Mental Mathematics Test 11* explains,

It is important to realise that, however carefully educational tests are constructed, they are accurate only to 'plus or minus' the confidence band. On another occasion, on a similar test, the same pupil is likely to achieve a different score.

(Clausen-May *et al.*, 1999, p.26)

For this test, the 90 per cent confidence band had been calculated as 'plus or minus 8' for a pupil with a standardised score of 100. In other words, we could be 90 per cent certain that the same pupil, on another, similar test, would score between 92 and 108. But if we recall that two-thirds of all the pupils in the standardisation sample scored between 85 and 115, then all we can really say is that the average pupil with a standardised score of 100 probably lies somewhere within the middle third. Equally, the confidence bands of Sally and Jenny, whose standardised scores are given above, clearly overlap. It is possible that their positions would be reversed next time they took a similar test.

Furthermore, the bands are only 90 per cent confidence bands, not 100 per cent. There is a ten per cent probability that a pupil's true score does not even fall within the 'plus or minus 8' band. As the *Teacher's Guide to the Mental Mathematics Test 11* explains,

since we are only 90% confident of the 'plus or minus 8' bands, out of a group of 30 pupils, we can expect three to have true scores that fall outside their confidence bands. However, we do not know which three pupils these are.

(Clausen-May *et al.*, 1999, p.26)

This is one reason why a test with a lot of marks can tell us more than one which has only a few. In a very short test, the confidence band for a pupil scoring half of the marks, say, might

cover most of the test. In that case we could have very little confidence in the result as an indicator of the pupil's performance.

### ***6e) Cut scores***

Sometimes all that is wanted from a test is a set of standardised scores, so that a teacher can see roughly how the pupils in a class compare with pupils nationally. But quite often, tests are used to allocate levels or grades, or to select pupils for some purpose. When this happens, a decision has to be taken relating to the *cut score* – the number of marks needed for a pupil to have passed the test or to have achieved a particular level or grade. In the latter case, there is usually an implicit assumption that the assessment is to some extent criterion referenced – that a pupil who gets 25 per cent of the marks, for example, can be expected to know this and that, while a pupil who gets 50 per cent will know this, that *and the other*.

Ian Schagen and Dougal Hutchison suggest in their analysis of the move from criterion referencing to mark-based cut scores in the early years of National Curriculum assessment that the link between Statements of Attainment (which represented specific criteria according to which levels could be awarded) and 'assigned Levels' was broken by the introduction of a mark-based system, so that

it is not clear to what extent such a system can be called 'criterion referenced'.

(Schagen and Hutchison, 1994, pp.211-21)

Steve Sizmur and Marian Sainsbury argue that the use of level descriptions, which replaced Statements of Attainment, offers a more sophisticated approach which may be sensitive to 'the underlying educational goals of the curriculum' (Sizmur and

Sainsbury, 1996, p.11). None the less, they point out that the complexity of the level descriptions, which are primarily intended for teacher assessment,

presents a considerable challenge to teachers, on the one hand, and to test developers, on the other.

(Sizmur and Sainsbury, 1998, p.192)

The level descriptions cannot be translated directly into a specific cut score in a test. Rather, the process of selecting a cut score involves a range of statistical considerations, for example to ensure that the right proportion of pupils attain each level or grade. In addition, teacher judgement may be taken into account, and the cut scores set at the point which matches the minimum score which, in the teachers' experience, a pupil who is working at each level should obtain. Thus it may take into account both a 'norm-referencing' and a 'criterion-referencing' approach. The whole process is complex, and the details cannot be gone into here – but test users should at least be aware of the difference that a small change in the cut scores may make.

If the cut score for a test carrying 100 marks is 50, then it is obvious that a pupil who scores 49, say, is much closer to a pupil who scores 50 than to one who scores only 20. The fact that the pupils scoring 49 and 20 have both 'failed', while the pupil scoring 50 has 'passed', is incidental. However, it is less obvious what effect changing the cut score will have on the outcome of the test. This will be dependent on a number of factors, including the total number of marks in the test and the nature of the distribution of the pupils' results – whether they are well spread out, or tightly bunched around the cut score. But to take a fairly straightforward example, we may consider a test carrying 100 marks, for which formal trialling with a large sample of pupils gives a normal distribution of scores with a mean of 50 and a standard deviation of 7.5. In that case, if the cut score were set at 50, half the pupils in the sample would achieve it, and

'pass', while the other half would 'fail'. But a change of just one mark in the cut score, to 49, would increase the percentage of pupils who passed by five per cent, to 55 per cent. Similarly, if the cut score were raised by one mark, the percentage of pupils who passed would be reduced by five per cent. Thus a change in the cut score of just one mark (in the test carrying 100 marks, with a normal distribution) could be expected to raise or lower the percentage of pupils achieving the relevant grade or level by five per cent when the final version of the test is taken. The smaller the number of marks in the test – and thus the shorter the test – the greater the effect of such a change in the cut score is likely to be.

### ***Statistics for Test Users: Key Points***

*Content validity* relates to the way in which tests actually do assess what they purport to assess. *Predictive* and *concurrent validity* are statistical concepts.

The *reliability* of a test is also statistically defined, in various different ways.

*Standardised scores* enable us to compare the results of a single pupil with those of a large, representative sample, and to compare scores on different tests.

A pupil's *true score* is what they would have got if all random variations resulting from *measurement error* could have been avoided.

There is a 90 per cent probability that a pupil's *true score* will fall within the *90 per cent confidence band*. However, the *true score* of one in ten pupils is likely to fall outside this band.

If a test is to be used to select pupils for some purpose, or to allocate levels, then a *cut score* must be identified. A small

change in the cut score may significantly affect the number of pupils who achieve it.

## 7 Statistics for Test Developers

*While the previous chapter focused on some more generally relevant statistical concepts, in this chapter, attention is paid to aspects of statistics which may be of greater interest to those who have overall responsibility for the development of a formal test. These include **samples**, which must be representative and of an adequate size to ensure that any statistical data that are collected are meaningful. The data are likely to include the **facility** of each question, which indicates its level of difficulty, and also the **point biserial correlation coefficient**, which gives a measure of its discrimination.*

### **7a) Samples**

Informal trialling may be carried out on a small sample, with just a handful of pupils, but one of the main purposes of formal trialling, whether of individual questions or of a whole test, is to collect sound statistical data relating to the questions being trialled. Obtaining high-quality statistical data depends on having *samples* of an adequate size. In an individual question trial, a sample of 350 pupils may be adequate. However, for a standardisation trial, a larger sample is required. If this involves an age standardisation, then it is the number of pupils in each age band which must be considered.

It is also important to ensure that the sample is representative – that it really does represent the population of pupils for whom the test is designed. For example, if the test is to be used throughout the United Kingdom, it should not be trialled only in Southern England. However, a test may also be trialled with a specifically selected biased sample. For example, a group of schools which have greater than average numbers of pupils with English as an additional language might provide a special subsample, to enable the statistics for these pupils to be



compared with those of the mainstream sample. The reliability of such statistics will again depend upon the number of pupils in the subsample.

If the test (or set of tests) is designed to distinguish performance at a range of levels, then a different sort of sample may be used, composed of roughly the same number of pupils at each level. Instead of the bell-shaped curve of a normal distribution, this will give a more or less straight horizontal line, called a rectangular distribution. For example, the great majority of pupils who take a Key Stage 3 national test will achieve levels 4, 5 or 6 – so if the tests were trialled with a representative sample of such pupils, there would be plenty of useful data on questions set at these levels, but relatively little on questions set at level 3 or level 7. For this reason, schools might be asked to identify pupils who are working at the extreme ends of the attainment range, and a disproportionate number of these pupils might be selected for the trialling sample.

### ***7b) Facilities***

The first set of statistics which are likely to catch the test developer's attention are the *facilities*. These relate to each individual mark in the test. It is essentially quite a straightforward idea. The facility of a mark is normally the percentage of the pupils who took the test who were awarded the mark. The facilities of pupils in different subgroups may also be found, such as the percentage of pupils with different Teacher Assessment levels, or with English as an additional language, who got the mark. In some situations the test developer may be more interested in the percentage of those pupils who attempted the question who were awarded the mark – so, for example, an easy question at the end of the test might have a low facility overall, because many pupils did not reach it, but a much higher

facility if only those pupils who attempted that question and the one following it are considered.

Any test needs a range of facilities, particularly if it is to be used to sort pupils into a number of different groups. It is good practice to put the 'easier' questions (the ones which greater numbers of pupils get right) at the beginning of the test, to make sure that pupils do not miss out on 'easy' marks because they get held up on a hard question early in the paper, and do not reach the end. Similarly, if a question has several parts, and carries a number of marks, then the earlier parts of the question should not be harder than the later, so they should not have lower facilities.

If the questions in a paper cover a range of levels of difficulty, and pupils have a given amount of time in which to do as much as they can, then it is to be expected that some pupils will not reach the end. If the paper is untimed, or if a very generous amount of time is available, then they may all 'complete' it – in the sense of being able to have a go at all the questions. However, the first of these arrangements may be difficult to manage, while the second may leave the majority of pupils with nothing to do for a long period. It also prevents the use of questions which reward efficient strategies and penalise inefficient ones.

When the trialled questions are marked, those which pupils have not attempted may be coded separately, to distinguish them from those which pupils attempted but got wrong. This allows the test developer to find what proportion of the pupils did not even attempt each question. If a high proportion of the pupils have omitted the final questions on the paper, then this could indicate that a lot of pupils ran out of time, unless the questions were so demanding that most pupils decided not to attempt them.

But although a range of questions is needed in a test, it is not worth including many which are either much too hard or much

too easy. A question which nearly all the pupils can do, or nearly all cannot do, will not help us to distinguish between the great majority of pupils. As a rule of thumb, for most questions in a test, not less than 40 per cent and not more than 70 per cent of the pupils for whom the question is designed should get the mark, while most marks should be gained by between 50 per cent and 60 per cent of the target group. This rule is not hard and fast: if a test were designed for the top ten per cent of 14-year-olds, for example, then if it were trialled by a representative sample of the whole age group an average facility of about ten per cent would be appropriate. On the other hand, if such a test were trialled with a specially selected group, say of pupils in selective secondary schools in areas where 20 per cent of all pupils got into to such schools, then one would again expect facilities of about 50 per cent – so about half of the 'top 20 per cent' would be getting the mark.

Furthermore, a test intended for mainstream pupils, trialled on a representative sample, might start with one or two questions with facilities of 80 per cent or more, or a question might have a high facility part at the beginning: these serve as 'ramps' to get the pupils in to the more demanding questions later on. There might also be an argument for putting a hard question with a low facility at the end of the test, to act as a challenge to the highest-achieving pupils, and to serve as a stimulus for class discussion later. But very easy or very hard questions are not efficient assessment tools: they do not tell us much about the pupils as a whole, except that nearly all of them can, or cannot, answer the question correctly. They are not good discriminators... and that brings us on to the next useful statistic, known amongst test developers as the *point biserial correlation coefficient*.

### ***7c) Point biserial correlation coefficients***

The *point biserial correlation coefficient* is a measure of discrimination. Like the facility, it is found for each individual mark in the test, but it also relates to the test as a whole. It connects the pupils' scores on each question with their overall test scores. The point biserial correlation coefficient answers the questions *Did most pupils who got this question right also get most of the rest of the test right? And did most of the pupils who got it wrong get most of the rest of the test wrong?* A perfect correlation would give a coefficient of 1; no correlation at all would give a coefficient of 0, while a negative coefficient would indicate that the question was often answered correctly by low-achieving pupils, but high-achieving pupils tended to get it wrong. So the higher the correlation coefficient, the better – all else being equal.

However, a low point biserial correlation coefficient is not necessarily a reason to drop a question from a test. This is because the magnitude of the coefficients in a test depend, in part, on whether all of the questions are actually assessing the same sorts of ability. If a question has a high correlation coefficient, then it is likely to be assessing what most of the other questions in the test are assessing – so pupils who do well with this question tend to do well in the rest of the test, while pupils who cannot answer this question correctly tend to do badly. On the other hand, a question with a low point biserial correlation coefficient is not assessing the same things as the other questions in the test. It may not be assessing anything at all – or nothing of interest to the test developer. Alternatively, it may be assessing other aspects of the curriculum which should be covered in the test. (Lord and Novick, 1968, p.381).

If a test is designed to cover a variety of different skills and abilities, then performance in one may not necessarily correlate closely with performance in another. For example, a mathematics test might include a lot of questions which require

pupils to carry out computations, but also a smaller number of questions which assess their spatial ability. On the whole, any one pupil will tend to do either well or badly on most of the computational questions, so these questions will have high correlation coefficients. The pupils who do well with the computations, however, may do quite badly with the questions which require them to use their spatial ability, while the pupils who could not do the computations may do better with a shape-based puzzle. So, for example, pupils who are good at computations such as:

$$\begin{array}{r} 367 \\ \times 46 \\ \hline \end{array}$$

are also likely to be able to answer the question,

John has **420 seeds**.



He wants to plant the seeds in **trays**.

Each tray can hold **6 rows of 8** seeds.

**How many trays** does John need?

These pupils may also be successful with questions like:

There are **32 tins** of beans on a shelf.

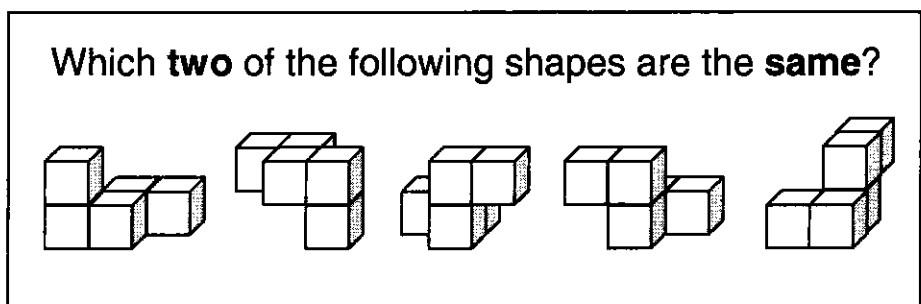
Each tin weighs **450 grams**.

What is the **total weight** of the tins?

Give your answer in **kilograms**.

All of these questions require pupils to multiply or divide, and a pupil who can answer one is likely to be able to answer the others. The questions are all assessing the same sorts of skills – so they are likely to have high point biserial correlation coefficients.

On the other hand, many of the pupils who could multiply and divide successfully might not be able to answer a question such as:








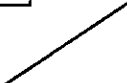
Furthermore, a number of the pupils who could not work out the answers to the first three questions might well be able to rotate shapes in their minds – to orientate them – effectively, and so gain the mark for the last question. This will lead to a low point biserial correlation coefficient for the question, which assesses spatial ability. If all the pupils who did really well with the numerical questions failed to answer the spatial question correctly, while all those who did badly on the numerical questions got the spatial question right, then the coefficient could even be negative.

This poor correlation between performance on questions which assess computational skills and those which measure spatial ability sometimes leads to the suggestion that questions on shape and space have 'poor discrimination' – but of course, this is true only in the context of a heavily Number-oriented curriculum and assessment structure (Clausen-May and Smith, 1998 and Clausen-May and Lord, 2000). To say that the question about

the rotated shapes has poor discrimination when it is placed in a test with a lot of questions requiring numerical computations means only that it is not a good measure of the pupils' ability to do the computations. This is hardly surprising – but none the less there may be pressure to cut out the 'erratic' spatial ability question. A substitute which assesses vocabulary may be found – something like:

Match each **shape** to its **name**.

The first is done for you.

					
					
	circle	triangle	square	pentagon	hexagon

This question, which is heavily dependent on linguistic ability but has little to do with spatial ability, is likely to correlate better with the computational questions – pupils who can do sums can often also learn new words, even if they have very poor spatial ability. Yet the question appears to address the same curriculum area as the one about the rotated shapes – so 'curriculum coverage' may be achieved. In this way, a test which already has many more questions assessing numerical than spatial skills becomes even more unbalanced. Such a test is clearly biased against spatial thinkers, but they may be yet further penalised by having the handful of questions which they could have answered correctly removed, and replaced by linguistically based questions. As Lord and Novick argue,

maximising reliability may sometimes be an undesirable goal. For example, a subset of factual

items in an achievement test may yield a more reliable score than the total set of items. This can happen, for example, if the other items involve such hard-to-measure but important traits as reasoning ability and creative thinking. Validity is of prime importance in such a case; one would not wish to increase reliability by discarding items if this decreased validity.

(Lord and Novick, 1968, p.344)

Thus the test should be considered as a whole, in the context of the wider curriculum, rather than as a collection of isolated question statistics.

### ***Statistics for Test Developers: Key Points***

*Samples* must be representative of the population for which the test is designed, and large enough to produce meaningful statistics.

The *facility* of a question is the percentage of pupils taking the test who answered the question correctly.

The *point biserial correlation coefficient* of a question measures its discrimination. If most of the pupils who got the question correct did well in the rest of the test, and most of those who got it wrong did badly, then the coefficient will be high.

If the test questions are all measuring the same thing, then the *point biserial correlation coefficients* are likely to be high. If a question has a low coefficient, then it may be measuring something different.



## 8 Looking Ahead

*Crystal ball gazing is fun – and risky. It is easy to see with hindsight how changes were bound to affect our lives, but it is not so easy to foresee changes. None the less, a current book on test development cannot, in all conscience, totally ignore the impact of ICT (information and communications technology).*

*This chapter starts by considering some of the ways in which the rapid increase in the use of ICT could affect aspects of the school curriculum in general. Some recent applications of ICT to test development, such as the use of **OMR** (optical mark reader) and **OCR** (optical character recognition) sheets, and the compilation of **item banks**, are then discussed. The chapter concludes with a consideration of some more radical possibilities, including the exploitation of the facilities of ICT to develop qualitatively different types of question in more flexible, **computer-adaptive tests**.*

*Anyone involved with ICT-based assessment, whether as a test developer, a question writer or reviewer, or a test user, is likely to face many of the issues raised in this final chapter.*

### **8a) Information and communications technology**

At the beginning of the third millennium, reading the printed word and writing with a pen or pencil on paper is still at the very heart of our education system. Pupils who, for whatever reason, cannot read and write are likely to fall behind in every area of the school curriculum. They cannot access much of the material to be learnt, and they cannot communicate effectively with their teachers.

The reliance on print for teaching materials is evident in a wide range of school subjects. The works of Shakespeare are often studied and assessed as though his overall aim was to produce

printed books, not performances. Reading about, say, the classification of rocks and minerals, or the effect of friction on a moving object, may replace direct observation and experiment as a scientific activity. Printed, two-dimensional shapes are studied in detail in the mathematics classroom, but solids are often largely overlooked.

But this cannot go on for ever. The ability to read may continue to be required for ready access to screen-based information, but the emphasis on writing will not be maintained. At present, pupils come under strong pressure to learn to write – they are formally assessed on their ability to form letters, and to join them up in the correct, prescribed fashion. But most professionals make very little use of such traditional writing skills. They may scribble the odd note, perhaps, but any serious work intended for others to read is likely to be in electronic form. This has obvious advantages – it is much more easily transferable, and may be more legible. The reader can work directly on the new material, and use it in a variety of ways, not just to read. It is also accessible to a wider range of users – so, for example, an increasing number of computers in common use can read prose aloud if that makes it easier for the user to understand.

It cannot be long now before good keyboard skills, or the ability to speak coherently into a computer with voice recognition, become more highly valued in the educational setting. Teachers may be uncertain which keyboard they should teach pupils to use – the traditional but inefficient and irrelevant QWERTY keyboard which we have inherited from a previous technology, or one of the more rational layouts which improve typing speed and accuracy. This might lead teachers to simply avoid the issue by opting to move straight to voice-controlled input. But in any case, pen and pencil skills will soon fall into relative disuse. In 30 years or so formal, compulsory tests of handwriting will be a thing of the past. Using a pen will become an optional craft – a

delightful and, for some children, valuable activity, but by no means essential for those who find it difficult and tedious.

Such a change will merely bring school practice, which often lags behind, into line with what is already working practice in nearly every office in the country. But one of the things which is holding us up is the fact that virtually all formal academic assessment is still dependent on writing skills. Even in areas such as mathematics or science, a pupil who cannot write has to have special assessment arrangements simply because they cannot write. They cannot show what they understand and can do unless they can do it with a pen on paper. So teachers may feel in duty bound to keep pupils immersed in a writing-heavy curriculum, in order to ensure that this increasingly anachronistic way of working is totally familiar and will not come as a shock when pupils are formally assessed.

But as keyboards and monitor screens replace pen and paper more and more as the main means of interpersonal communication outside the classroom, schools will begin to change. Assessment cannot block progress for ever. The insistence on handwritten responses to a mathematics or science test, for example, will be seen as disadvantaging some pupils on the basis of their failure to perform an irrelevant skill, and this will help to force a change. With time, most tests will be ICT administered and, in many cases, ICT marked.

### ***8b) The beginnings of change***

ICT is already used to some extent in the context of assessment. For example, *OMR* (*optical mark reader*) sheets, which were mentioned in section 3a), have been in use since the 1950s. These allow us to do away with expensive markers: pupils complete the sheets by shading in the box containing the code number of the answer they have selected to each question, and

the optical mark reader 'reads' these selections and works out how many are correct.

Unfortunately, the use of OMR sheets imposes severe restrictions on the types of question which can be asked: they must all be multiple-choice or multiple-response, with no closed-answer and certainly no open-response questions. On the other hand, if *OCR* (*optical character recognition*) sheets are used, or if the pupil responds to the questions directly through the keyboard, it may be possible to include some closed-answer questions. However, this can give rise to all the problems discussed in section 5d) which are associated with assuming that there is one clearly defined correct answer to even a simple question.

Another way in which ICT has been used in the context of assessment is in the creation and storage of what are called *item banks*. An item bank is a large collection of individual questions, often with one mark each, all of which have been formally trialled with a good sample of pupils to establish their facilities and discriminations. The questions may also have other information stored with them, such as the area of the curriculum which they address, or whether they involve the representation or interpretation of data.

The existence of a large item bank stored on a computer allows the test developer to select a number of questions to form a complete test which, at least in theory, will have certain specified properties. For example, the developer may require the test to have a given mean and standard deviation, to cover a specified range of topics, and to have a particular proportion of questions which require pupils to interpret or represent data. A program may be devised to search the item bank for a suitable selection of questions, perhaps offering several different combinations from which the developer can make a choice.

Unfortunately, at least with the technology that is currently available, this approach to test development is not likely to be holistic (see section 2a). If the questions in an item bank are atomised and disconnected, with one mark each, then a test composed of a selection made on primarily statistical grounds cannot be expected to form a coherent whole. It is unlikely to convey any sense of the interconnections between different aspects of the subject being assessed. Furthermore, although all the individual questions in an item bank are well trialled in order to establish their statistical properties, the particular set of questions which is selected to form a test on the basis of those statistics may never have been trialled together. Unless a formal trial (see section 5c) is undertaken after the selection, changes to facilities which result from changing the order of the questions in the newly created test, for example, have to be ignored when the statistical parameters are specified. These problems may be overcome in time, with banks of full-length questions and more extensive trialling of different combinations, but at present tests compiled from an item bank (whether computer driven or manually selected) are not likely to offer an integrated and coherent assessment of the subject.

So the use of OMR or OCR sheets, and of computer-administered and marked tests, make the process of marking tests easier, and cheaper. A computerised item bank enables a test with particular parameters to be compiled more quickly. All these uses of ICT may be considered more efficient than doing the job manually – but they all impose greater restrictions on the types of question and test that can be created. The test developer is forced to rely heavily on multiple-choice and multiple-response questions, and to produce tests which are composed of discrete, disconnected questions which tend to undermine the development of a coherent curriculum.

But while they all save money, OMR and OCR sheets, computer-administered and marked short-answer tests, and item banks are

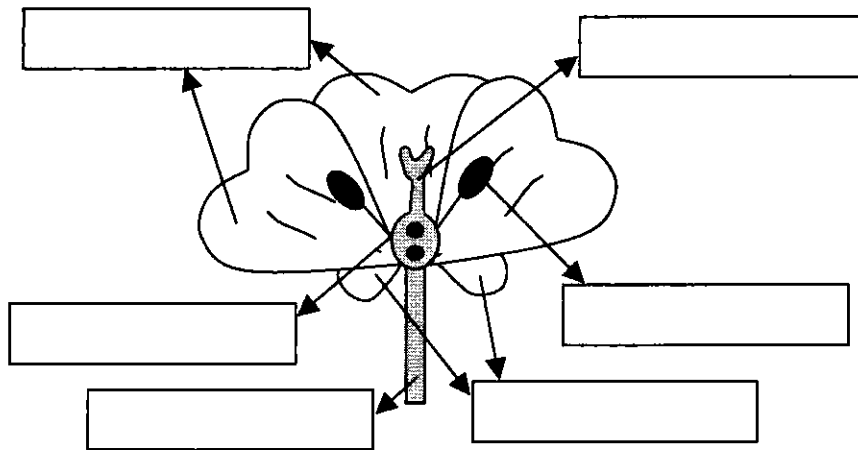
all, perhaps, just more QWERTYs. Just as the layout of the keyboard on most computers still reflects requirements which were imposed by the technology of the earliest typewriters, so these uses of ICT in the context of test development are restricted by the limitations of an earlier approach. Tests composed of multiple-choice, multiple-response and short-answer questions existed long before computers were in common use. Tests made up of questions drawn from a large pool of disconnected 'items' were not unknown. So in all these approaches to using ICT for test development, the new technology simply does more quickly and cheaply what could be done manually. The facilities of ICT are not really exploited to do something completely different.

### ***8c) What is to come?***

Since this book is produced in a print-based medium, it is difficult to give good examples of the sorts of question which could be developed to exploit the facilities offered by ICT. None the less, some possibilities may be suggested. These should reflect the changing curriculum as it adapts to a classroom in which pupils do most of their work on screen rather than in exercise books.

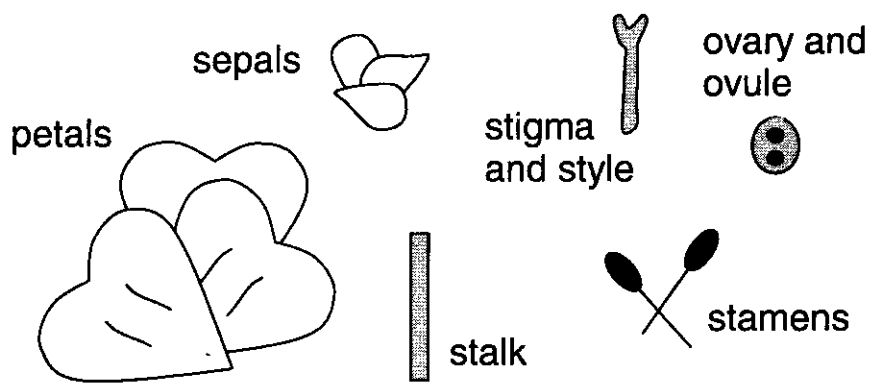
A simple use of ICT involves the exploitation of the ability to 'click and drag' on the screen. Anything may be clicked and dragged – pictures, words, numbers, or boxes containing a combination of these. For example, written primary science tests often include questions which require pupils to label the parts of a structure, such as a flower or a skeleton.

Label the parts of the flower in the diagram.

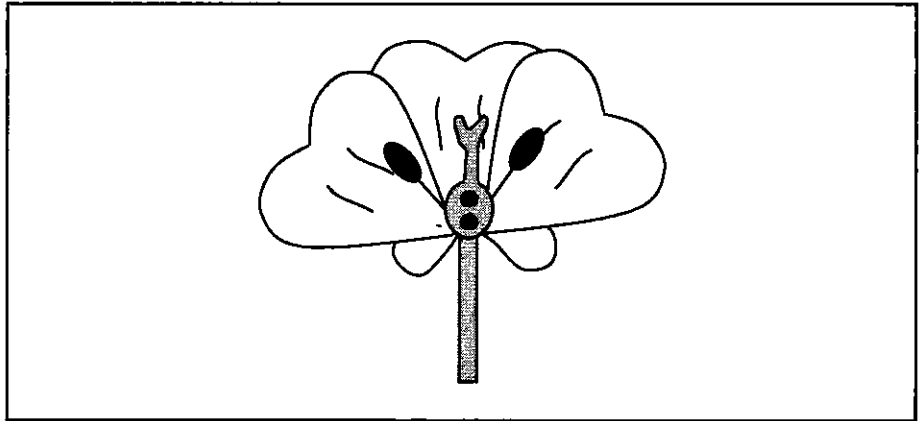


This question is essentially a vocabulary test: it assesses the pupils' memory for a set of terms, but it does not assess their understanding of the structure itself. An alternative approach might be to ask pupils to assemble the parts of a flower into a complete diagram.

Arrange these parts of a flower to make a diagram to show how they fit together.



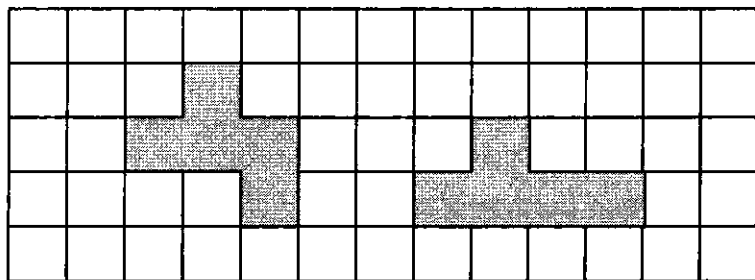
By clicking and dragging the parts, pupils could put together a diagram of the flower.



This question focuses on the pupils' understanding of the structure of the flower, rather than on the 'naming of parts'.

The use of movement on the screen allows a more practical approach to some areas of the curriculum to be assessed. For example, a mathematics question could start by demonstrating how two shapes may be rotated and flipped (reflected), then fitted together to fill a given frame. The screen would first show the two shapes separated.

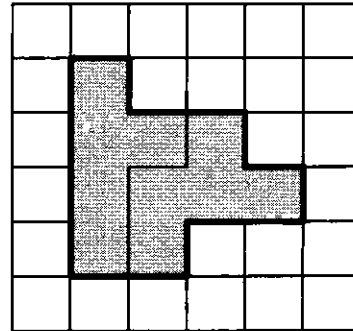
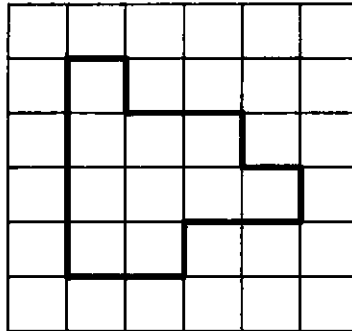
Look at the two shapes.



The pupils would be shown an example of the way in which the two shapes may be fitted together to fill a frame.

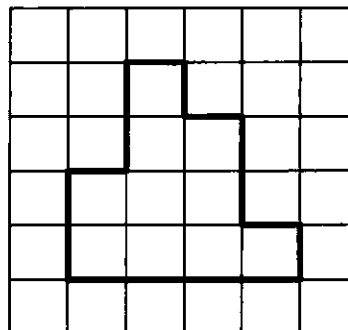


The two shapes can fit together to fill this frame.



The screen would show the two shapes moving into the frame. Then the pupils would be asked to fit the shapes into a different frame.

Fit the two shapes into this frame.



Pupils would be able to rotate and reflect the two shapes in order to place them into the new frame.

Other ways of harnessing the power of ICT to develop innovative types of question, rather than simply putting standard written questions on screen, need to be explored. Furthermore, with computer administration it would be possible to measure

the amount of time each pupil took to respond to a question, and to take account in the assessment of the difficulty they had in reaching a correct solution.

### ***8d) Computer-adaptive testing***

In a conventional, paper-and-pencil based test, every pupil answers exactly the same set of questions, usually in the same order. However, this is not always very efficient, especially for pupils working well above or below the norm. Very able pupils have to waste time answering what may, for them, be trivial questions. Perhaps more seriously, less able pupils may be exposed to a very depressing experience which undermines what little confidence they may have had, as they sit and fail to make progress with question after question in the test.

An alternative strategy is to start by offering pupils a question with several parts, or a small set of shorter questions, with facilities near or below the middle of the range covered by the test. Pupils who have no difficulty at all with this may then be moved straight on to some significantly harder questions: pupils who appear to struggle, perhaps by accepting supportive feedback which may be offered on screen, should be given questions pitched at a lower level. In this way, the test quickly homes in on a set of questions which are appropriate to the particular pupil who is taking it. Data which are more detailed and reliable may be collected with fewer questions than is possible using a conventional test.

Another ambitious way in which ICT may be used in a test is by adapting questions to the individual pupil, offering feedback and prompts where appropriate. A question may be presented with no support, and the pupil invited to respond. If the pupil cannot respond, or responds incorrectly, however, then help may be given. Clearly, a pupil who needs a lot of help is working at a

lower level than one who needs none, and might be awarded fewer marks – but at least this approach would give all pupils a better chance of understanding what the question was about, even if they could not answer it by themselves. The marking structure could also take account of the time taken by pupils to respond to questions, so that a more efficient method would gain more credit than a slower, less efficient one.

For example, the question about fitting two shapes into a frame which was given above could be asked first without the facility to move the shapes. Pupils who could visualise the movement would be able to draw in the line dividing the two shapes, using the mouse and following the square background grid. However, those who needed the extra support could be offered the opportunity to virtually manipulate the shapes.

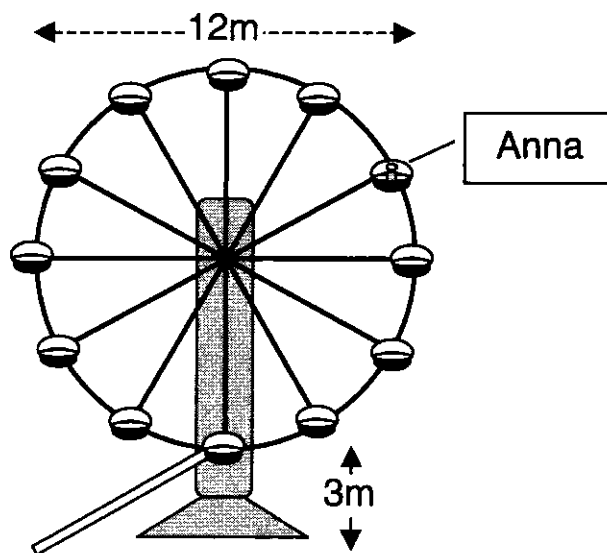
Again, for pupils working at a higher level in mathematics, the ability to analyse a situation and sketch a relevant diagram – to model the situation mathematically – is an important skill, and one that is worth assessing. However, some pupils may fail this initial hurdle, and be unable even to attempt to answer the question. Trialling two versions, with and without a diagram, may show that only by providing the diagram will the required question facilities be achieved. But the decision to include the diagram is usually taken with regret, as it is seen as providing too much scaffolding, and reducing the value of the question as an assessment of pupils' ability to mathematise effectively.

With ICT, however, the decision on whether or not to provide the mathematical diagram may be left until the pupil is attempting the question. For example, the following question does include a picture which encompasses many of the features of the required diagram, but not the diagram itself:

Anna is on a Ferris wheel which has 12 equally spaced chairs.

The wheel has a diameter of 12 metres.

The bottom of the wheel is 3 metres from the ground.

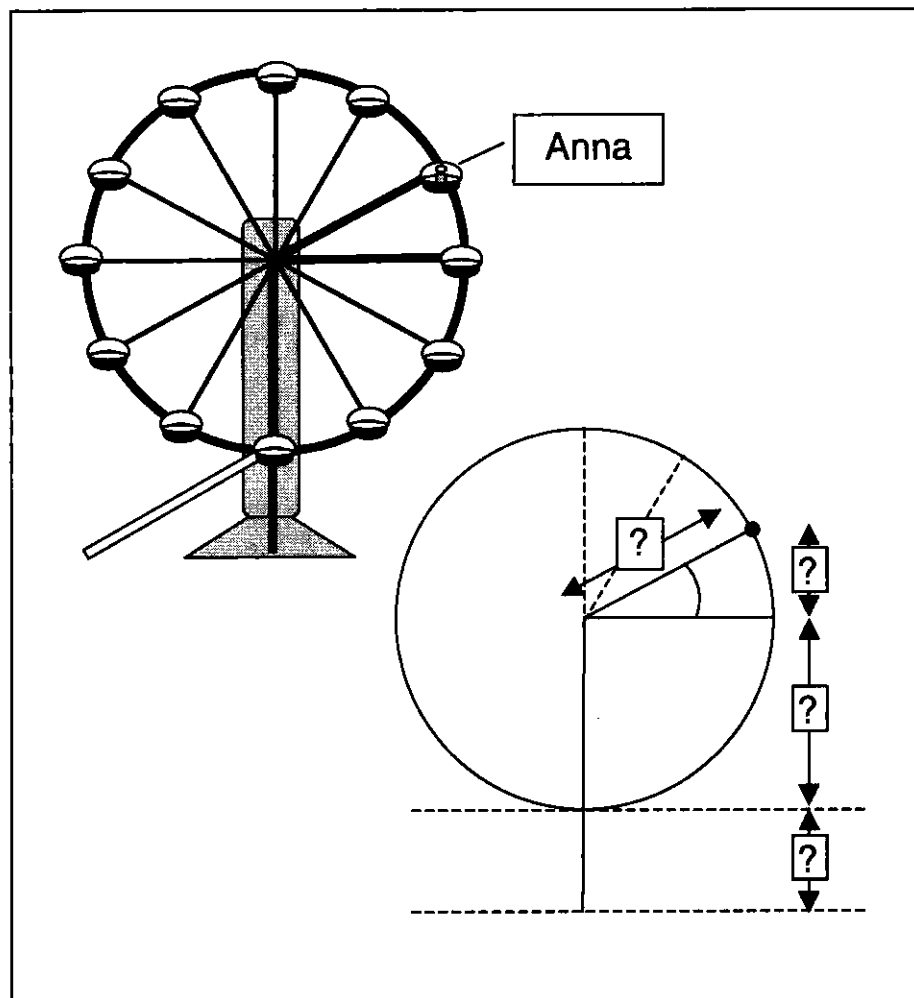


The wheel has stopped in the position shown in the picture.

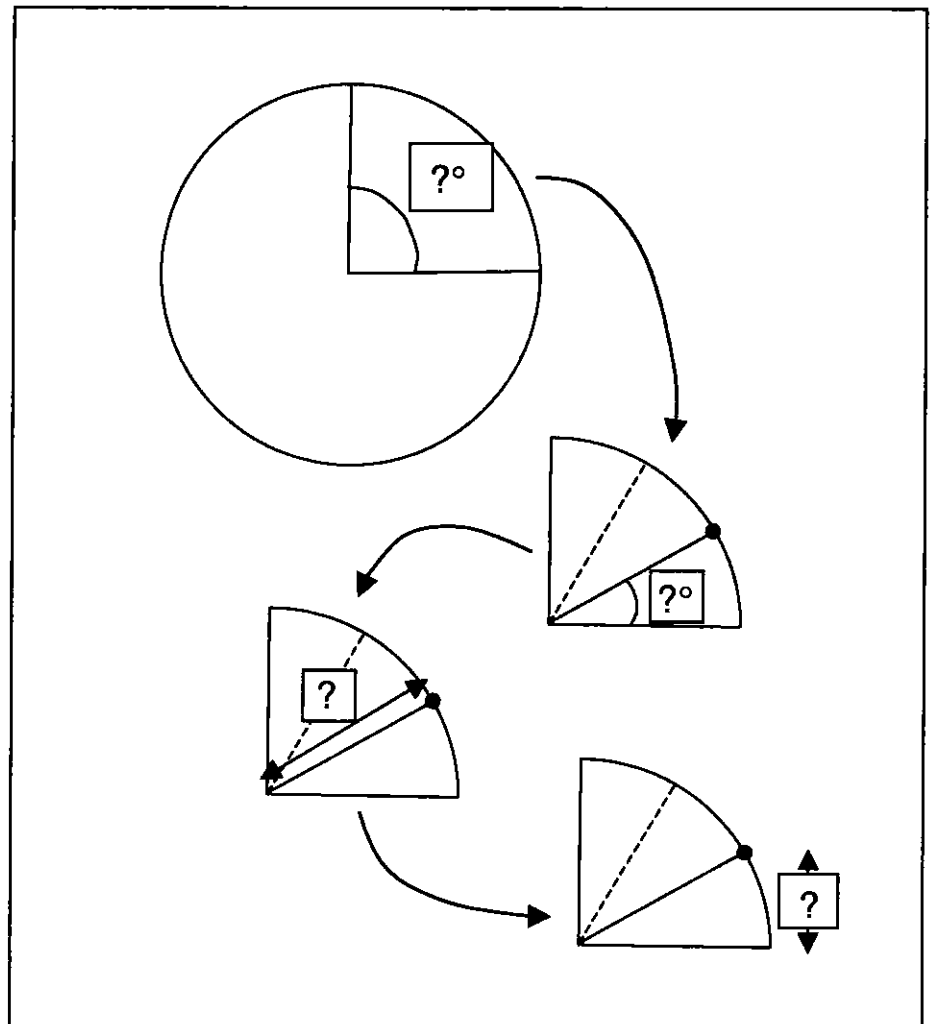
How high off the ground is Anna?

The most able pupils will be able to sketch their own diagrams, selecting the relevant information and ignoring the irrelevant. Other pupils, however, might be guided through the question in several stages.

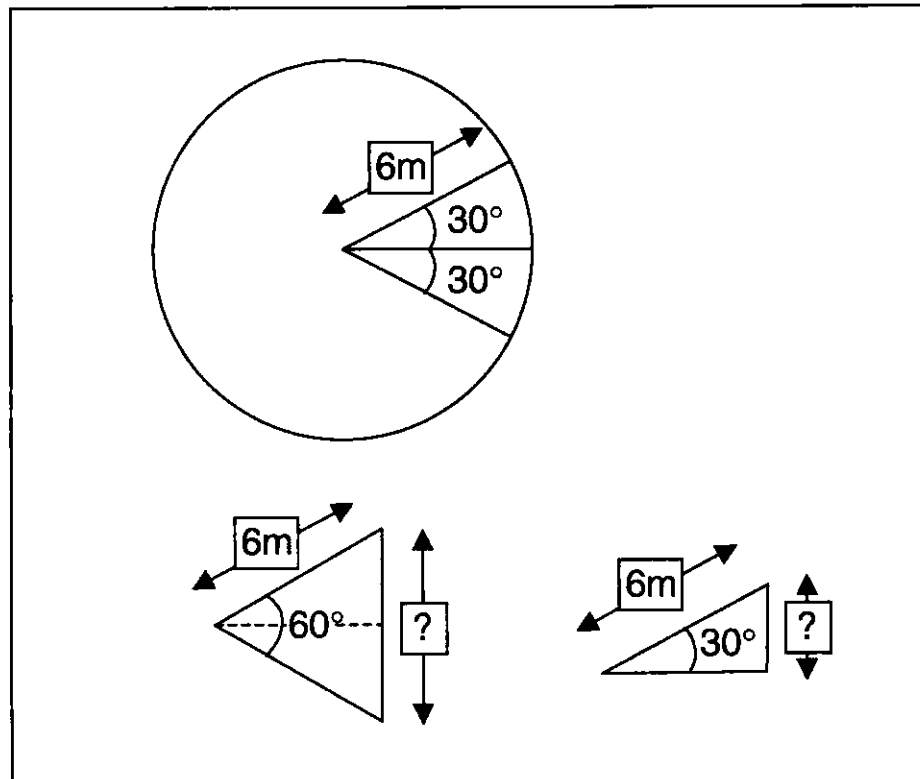
The computer could start by offering an unlabelled diagram. This could be superimposed on the picture, and then moved away so it could be seen more clearly.



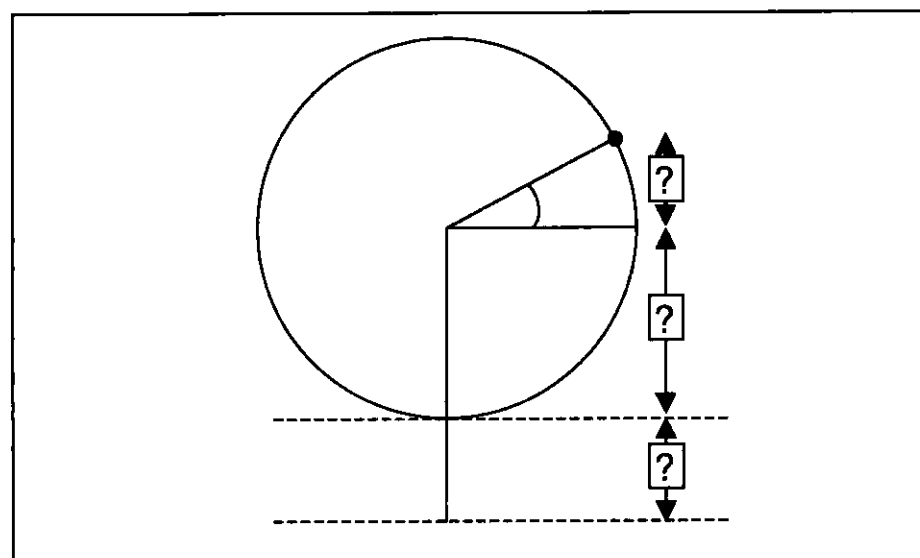
The pupil would be asked to fill in the empty boxes with the relevant figures. If this proved too difficult, then further prompts could be given. For example, the height above the horizontal could be prompted with series of diagrams:



Some pupils might be able to proceed from here unaided. If the height above the horizontal still proved difficult to find, however, then further help could be offered:



Now the pupil could, if necessary, be prompted to return to the first diagram, and to fill in all the distances needed to find the total height of the chair off the ground.



This fairly detailed description of the different levels of feedback and support that could be given, tailored to the needs of different pupils, is intended to indicate how this approach might be developed. The clear distinction between teaching and assessment is blurred: a pupil who is led through the question is certainly being taught, but information relevant to the assessment of their level of understanding, and of their learning style, is also being gathered. Thus the functions of summative and of formative assessment, and of teaching and learning, are brought together.

Furthermore, this one question is effectively pitched at a range of levels. It offers a useful challenge to pupils who can structure their own approach and pull out the essential features to construct an appropriate diagram. But it also offers support and guidance through the question for those who need it.

But there is just one word of warning. As Randy Bennet, in a report from the Policy Information Center, in Princeton, New Jersey, points out,

Obviously, human judges will... retain purview over those tasks which are not computer deliverable – some important tasks won't soon be amenable to assessment in this medium (e.g. in the performing arts).

(Bennett, 1998, p.9)

In other words, even when ICT-based assessment becomes more sophisticated, there will still be some activities which cannot be assessed through ICT. There is, perhaps, a danger that these activities will be downgraded as a result. Aspects of the curriculum which depend upon teacher assessment are commonly regarded with suspicion – so the outcomes of coursework in the GCSE assessment structure, for example, are felt to be less dependable than written test results. In the future, it could be that ICT-based test results will be valued, and anything else will be taken less seriously.



Selecting sets of questions which are pitched at a level which is appropriate to individual pupils, and adapting questions to meet their individual needs, are both aspects of *computer-adaptive* testing. At present computer-adaptive tests which are composed of multiple-choice or multiple-response questions are being developed, but the range of question types used could be broadened in the future. This approach could have a number of advantages.

Computer-adaptive tests are not tied down to rigid, standardised assessment procedures, in which no account is taken of the special assessment needs of particular pupils. Computer-adaptive testing is adaptable: that is self-evident. As it becomes accepted as the norm, the idea that everybody should be assessed in exactly the same way with exactly the same test will become obsolete. Concerns about special arrangements will fade. If different pupils are taking different tests in any case, then a few more differences will be seen as trivial.

Again, computer-adaptive questions offer pupils different paths through a question, dependent upon the level at which they are working, and perhaps on their preferred thinking and learning style. This will enable test developers to create questions which probe more deeply into the pupils' understanding of the principles which underlie their knowledge. For example, the question about the Ferris wheel, with its accompanying prompts, could be used to identify pupils who can, and those who cannot, mathematise the situation and draw appropriate diagrams for themselves. The distinctions between testing, teaching and learning may become blurred – but the test itself will be of greater value, and will focus on the pupils' level of understanding. Every assessment will be formative, and, with computer-adaptive testing, one of the things that it will help to form will be the test itself.

### ***Looking Ahead: Key Points***

The school curriculum is still very dependent upon writing, but as access to ICT increases, keyboard and dictation skills will replace writing in schools as elsewhere.

ICT is already used in test development, for example with the use of *OMR* (*optical mark reader*) and *OCR* (*optical character recognition*) sheets, and for the creation of *item banks*.

Many of the current uses of ICT impose severe restrictions on the types of question and test that can be developed. They merely enable test developers to do more quickly and cheaply what could be done without ICT.

In the future, the use of ICT could enable the development of qualitatively different types of question.

ICT allows the use of *computer-adaptive testing*, in which both the test as a whole and individual questions within it may be adapted to suit the particular pupil taking the test.

## **Appendix:**

### **Writing Multiple-choice and Multiple-response Questions**

#### ***Introduction***

In the first four chapters of this book, some aspects of the process of writing test questions were discussed. It was argued that it is very difficult to write closed or multiple-choice questions which can distinguish between the learner who has merely procedural knowledge and one who has a higher level of conceptual understanding.

None the less, such questions are often used in tests. They have a number of advantages: they are straightforward to administer, and they are easy, and therefore relatively cheap, to mark. Postlethwaite reports that in one international study,

When there were only multiple-choice items the cost per pupil was \$5 but when open-ended items were introduced the cost went up to \$25 per pupil. Cost is therefore a major issue when deciding on item type.

(Postlethwaite, 1999, p.38)

The great difference is in the cost of marking closed and open questions – and this might have been exacerbated in this case by the need to consider pupils' possible responses in a range of different languages.

The general principles for writing good test questions apply as much to closed or multiple-choice as to open-response questions. The use of simple, direct sentences, careful layout, the appropriate use of boldening, and so on are as important here as elsewhere. However, there are a number of issues which are specific to the writing of multiple-choice or multiple-response questions. This appendix, bringing these points together, is based on a set of guidelines developed at the National Foundation for Educational Research for question writers working on a published multiple-choice test.

### ***Guidelines for writing multiple-choice and multiple-response questions***

A multiple-choice or multiple-response question consists of two parts – a *stem*, in which the problem is contained, and a range of *options*, one or more of which are correct. Incorrect options are called *distracters*. A stem can be either a direct question or an incomplete statement which can be completed correctly by one of the options.

Each question must have at least four options. In a multiple-choice question, there is one correct answer with at least three distracters. A multiple-response question can have more than one correct answer among the options. The pupil may be told how many there are, and is required to identify all of them to gain the mark.

### ***Language***

The questions need to be accessible to a wide range of pupils, who may have very different language and reading abilities. The language used must be clear, simple and concise. Technical terms should be used only when the terms themselves are being tested; otherwise such terms may make questions inappropriately inaccessible.

- Use clear and simple language and avoid ambiguous terms.
- Use the active, present tense in stems and options.

### ***General guidelines***

- Avoid using double negatives in either the stem or an option.
- Avoid unnecessarily difficult or technical language where possible.
- Aim for independence among questions. This means that, as far as possible, the options to one question should not provide the answers to another question.
- Use negatives only occasionally in a question. Negatives such as

'not', 'never', etc. may be made to stand out by using capitals or bold. However, the emboldened wording of the question as a whole should be considered, to ensure that a pupil who reads only the bold will be able to pick out all the key information in the question, as described in section 4c).

- Use two short sentences, rather than one long one, if this will make the question clearer.

### ***Guidelines for writing the stem***

- Do not make stems too long. A maximum of 50 words should be used, using a maximum of four sentences. Ideally sentences should be less than 15 words long.
- Avoid the use of conditionals, for example 'if ...'.
- Sometimes it may be necessary to ask which is the 'best' answer rather than the correct answer. Generally though, questions should be written which contain an answer which is clearly correct. (Multiple-response questions, however, may have more than one answer which is correct.)
- Phrase the stem so that a knowledgeable pupil can give the answer without seeing the options. The stem should not include general instructions, for example 'tick the answer below that ....'. It should contain a specific scenario.
- Include as much of the problem as possible in the stem, so that the options can be kept short.

### ***Guidelines for writing options***

- Keep answer options short and concise – a maximum of 15 words should be sufficient for most options.
- Make all options approximately equal in length, and parallel in grammatical structure and general appearance.

- Ensure that each option follows logically and grammatically from the stem. The correct answer must not be grammatically different from the other options.
- Avoid the use of 'none of the above'.
- Avoid the use of 'all of these'.
- Make sure that the answer does not contain careless clues which identify it as being correct.
- Ensure that the distracters are clearly incorrect, but could appear plausible.
- If you use options which form a pair, for example by stating the opposite of each other, then make sure that the remaining two options also form a pair.
- Do not repeat words or phrases from the stem in the options.
- You may sometimes be able to make a question more difficult by creating options which are very similar to each other.

## References

- BENNETT, R.E. (1998). *Reinventing Assessment: a Policy Information Perspective*. Princeton, NJ: Education Testing Service.
- BLACK, P. (1998). *Testing: Friend or Foe? The Theory and Practice of Assessment and Testing*. London: Falmer Press.
- BRITISH ASSOCIATION OF TEACHERS OF THE DEAF and NATIONAL ASSOCIATION FOR TERTIARY EDUCATION FOR DEAF PEOPLE (n.d.). *Language of Examinations*. Beverley: BATOD and NATED.
- CLAUSEN-MAY, T. (1998). 'Common errors – common understanding: the development of KS3 mathematics tests for fourteen-year-olds.' Paper presented at the British Educational Research Association Annual Conference, University of York, 11 September.
- CLAUSEN-MAY, T., CLAYDON, H. and RUDDOCK, G. (1999). *Mental Mathematics 11* (Mental Mathematics 6–14 Test Series). Windsor: NFER-NELSON.
- CLAUSEN-MAY, T. and LORD, T. (2000). 'Thinking styles and formal assessment: spatial and numerical thinkers in the mathematics classroom', *Mathematics in School*, **29**, 3, 10-13.
- CLAUSEN-MAY, T. and SMITH, P. (Eds) (1998). *Spatial Ability: a Handbook for Teachers*. Slough: NFER.
- DAVIS, R.D. (1994). *The Gift of Dyslexia*. London: Souvenir Press.
- GREAT BRITAIN. DEPARTMENT FOR EDUCATION AND EMPLOYMENT (1999). *From Inclusion to Exclusion: a Report of the Disability Rights Task Force on Civil Rights for Disabled People*. London: DfEE.

- LORD, F.M. and NOVICK, M.R. (1968). *Statistical Theories of Mental Test Scores*. London: Addison-Wesley.
- MESSICK, S.J. (1989). 'Validity.' In: LINN, R.L. (Ed) *Educational Measurement*. Third edn. London: Collier Macmillan.
- MOBLEY, M. (1987). *Making Ourselves Clearer: Readability in the GCSE* (Secondary Examinations Council Working Paper 5). London: Secondary Examinations Council.
- POSTLETHWAITE, T.N. (1999). 'Overviews of issues in international achievement studies', *Oxford Studies in Comparative Education* (Special Issue: Comparing Standards Internationally: Research and Practice in Mathematics and Beyond), **9**, 1, 23–60.
- QUALIFICATIONS AND CURRICULUM AUTHORITY (1998). *Standards at Key Stage 3 Mathematics: Report on the 1998 National Curriculum Assessments for 14-year-olds. A Report for Headteachers, Heads of Department, Mathematics Teachers and Assessment Coordinators*. London: QCA.
- SCHAGEN, I. (1999). 'Testing, testing testing', *Managing Schools Today*, **8**, 4, 28–9.
- SCHAGEN, I. and HUTCHISON, D. (1994). 'Measuring the reliability of National Curriculum assessment', *Educational Research*, **36**, 3, 211–21.
- SIZMUR, S. and SAINSBURY, M. (1996). 'Criterion-referencing and level descriptions in National Curriculum assessment', *British Journal of Curriculum & Assessment*, **7**, 1, 9–11.
- SIZMUR, S. and SAINSBURY, M. (1998). 'Level descriptions in the National Curriculum: what kind of criterion referencing is this?' *Oxford Review of Education*, **24**, 2, 192.







## **An Approach to Test Development**

Tests are having an increasing impact on what goes on in schools.

A good test can support and inform teachers in their work. It will encourage good teaching practice, with a coherent, holistic approach to the curriculum. It will focus on pupils' understanding, rather than on their memory. And it will be accessible, as it stands, to pupils with a wide range of special educational and assessment needs.

This book explains some of the nuts and bolts of sound test development. Written by a teacher and test developer, it will be of interest to anyone who wants to

- choose a test to use in the classroom;
- interpret test results;
- write test questions;
- manage the development of a large-scale testing project.

Using straightforward, readable language, **An Approach to Test Development** offers an insight into the sometimes puzzling world of tests and testing.

ISBN 0 7005 3021 5  
£10.00