

Designing Assessment of Performance in Mathematics

Hugh Burkhardt and Malcolm Swan ¹
MARS: Mathematics Assessment Resource Service
Shell Centre Team, University of Nottingham

Summary

The effective implementation of intended curricula that emphasise problem solving processes requires high-stakes tests that will recognize and reward these aspects of performance across a range of contexts and content. In this paper we discuss the challenge of designing such tests, a set of principles for doing so well, and strategies and tactics for turning those principles into tasks and tests that will work well in practice. While the context is England, the issues raised have wider relevance.

1. Introduction

Everyone concerned with education recognizes the importance of assessment – but most would like to minimise its role. Parents accept tests as necessary hurdles on the route to valuable qualifications for their children – but are concerned at the pressure and the consequences of failure. Many teachers accept the importance of tests – but also feel threatened and, believing that they “know” their students capabilities, see the tests as a disruption of teaching and learning. Politicians see tests as the key to accountability and the way to prove the success of their initiatives – but want to minimize the cost and the pressures on them that complex tests generate, through appeals against scoring, for example. All would like the tests to be “fair”, objective, and easy to understand. These very different motivations have led to high-stakes tests of Mathematics in England that assess only some elements of mathematical performance, mainly concepts and skills tested separately.

The revised national curriculum for the age range 11-16 takes a much broader, more holistic view of performance in mathematics, in line with high international standards. It focuses on developing the “Key Concepts” and “Key Processes” below across familiar content areas: Number and Algebra; Geometry and Measures; Statistics.

Key concepts: <ul style="list-style-type: none">• Competence• Creativity• Applications and implications of mathematics• Critical understanding	Key processes: <ul style="list-style-type: none">• Representing• Analysing (reasoning)• Analysing (procedures)• Interpreting and evaluating• Communicating and reflecting
--	--

To achieve this it demands curriculum opportunities to: develop confidence in an increasing range of methods; work on more challenging mixes of contexts and mathematics; work on open and closed tasks in real and abstract contexts; tackle problems from other subjects and from outside school; link different concepts, processes and techniques; work collaboratively and independently; select from a range of resources, including ICT.

The specification of this “Programme of Study” embraces the essential ingredients for using mathematics effectively in problem solving in the outside world or, indeed, within

¹ A revised version of this draft paper will be submitted to *Educational Designer*, the forthcoming e-journal of ISDDE. We appreciate the opportunity to discuss some of the issues at ISDDE08.

mathematics itself(see e.g. Schoenfeld 1985, 2007, Blum et al 2007, Burkhardt with Pollak 2006, Burkhardt and Bell 2007):

- knowledge and skills,
- strategies and tactics;
- metacognition;
- attitude .

This paper is concerned with how assessment, particularly high-stakes tests, can be aligned with these performance goals.

Here we shall focus mainly on the core design challenge – the tasks that enable students *to show what they know, understand and can do* (Cockcroft 1982) and the scoring schemes that assign credit for the various aspects of performance. Tasks may be packaged into tests in many ways – a largely separate, sometimes contentious issue; we shall also outline a process for building balanced tests from tasks. Finally, we shall comment on the process of implementing improved assessment, turning principles into practice in a way that the system can absorb, without undermining the always-good intentions.

Our approach, as ever, is that of engineering research in education – “the design and development of tools and processes that are new or substantially improved” (RAE 2001, 2006). It reflects several decades of experience working with examining bodies, nationally and internationally.

2. The roles of assessment

Assessment of performance is an important part of learning in any field, whether it be playing a sport or a musical instrument, or doing mathematics. It provides feedback to the learner and teacher that should help guide future study and, from time to time when demanded, summative feedback for accountability and other purposes. To fulfil its roles, assessment needs to be *well-aligned* – i.e. balanced across the various learning goals.

Why? Why are ‘simple tests’ of ‘basic skills’ not good enough? They seem so attractive to teachers, parents and politicians (though for different reasons) It is often argued that, though such tests only measure a small part of the range of performances we are interested in, the results “correlate well with richer measures.” Even if that were true, it is *not* a justification for narrow tests. Why?

Basically, in a target driven system where results have serious consequences, *what you test is what you get*. Since the tests purport to embody the targets society sets for education, this seems reasonable; but if the tests cover only a subset of the performance goals, it produces gross distortion of learning². WYTIWYG is regarded as obvious and inevitable by most teachers. It is regularly observed by those who inspect English school (see, e.g. OFSTED 2006). But it is too often ignored in assessment policy and provision – with the inevitable consequences.

To make progress, it must be accepted that high-stakes assessment plays *three* roles:

Role A: measures levels of performance across the range of task-types used.

Role B: exemplifies performance objectives in a clear form that teachers and students understand, and through this

Role C: determines the pattern of teaching and learning activities in most classrooms.

If the tests fail to reflect the learning goals *in a balanced way*, Roles B and C mean that classroom activities and learning outcomes will reflect that imbalance.

² An example may help show why balance across the performance goals is crucial. If, for reasons of economy and simplicity, it were decided to assess the decathlon on the basis of the 100 meter race alone, it would surely distort decathletes’ training programmes. This has happened in Mathematics where process aspects of performance are not currently assessed – or taught.

The importance of 'alignment' between assessment and the curriculum goals is widely recognised internationally, particularly where narrow tests are undermining the achievement of those goals. In its *Curriculum and Evaluation Standards*, the US National Council of Teachers of Mathematics stressed that "assessment practice should mirror the curriculum we want to develop; its goals, objectives, content and the desired instructional approaches", adding

"An assessment instrument that contains many computational items and relatively few problem-solving questions, for example, is poorly aligned with a curriculum that stresses problem solving and reasoning. Similarly, an assessment instrument highly aligned with a curriculum that emphasises the integration of mathematical knowledge must contain tasks that require such integration. And, for a curriculum that stresses mathematical power, assessment must contain tasks with non-unique solutions." (NCTM 1989 p 194-5).

These points are again emphasized in the NCTM 2001 revision, *Principles and Standards for School Mathematics*.

To summarise in rather more technical terms, the *implemented(or enacted) curriculum* will inevitably be close to the *tested curriculum*. If you wish to implement the *intended curriculum*, the tests must cover its goals in a balanced way. Ignoring Roles B or C undermines policy decisions; accepting their inevitability has profound implications for the design of high-stakes tests.

This can be an opportunity rather than, as at present, a problem. Both informal observation (e.g. with well-engineered course work) and research (e.g. Barnes, Clarke and Stephens 2000) have shown that well-designed assessment can be a uniquely powerful lever for forwarding large-scale improvement.

3. Performance goals in mathematics

From a strategic perspective, four kinds of tool are needed to enable a planned curriculum change to be implemented as intended – *standards, teaching materials, professional development support, and assessment* that all reflect the same range of goals. In this paper we seek to link the first and last of these.

QCA has now described the *standards* in the revised Programmes of Study. These are an *analytic description of the elements* of the intended domain of learning. The key processes: *representing* a practical situation in mathematical terms, *analysing* this model of the situation, *interpreting, evaluating* and *communicating* the results, together imply substantial change in the performance goals – and in classroom practice.

However, as they stand, the Programmes of Study do not define performance goals. For example, the processes could be regarded as independent, and assessed separately, or as elements in an integrated problem solving process. These are very different kinds of performance. This is not an academic issue; it was the decision to test the elements of performance separately through short items that undermined the intended performance goals of the 1989 National Curriculum in Mathematics and led to the current almost-process-free curriculum that the revised Programmes of Study aim to improve on.

The overarching aim of the revised Programmes of Study is to align mathematics curriculum and assessment with authentic examples of thinking with mathematics about problems in the outside world and in mathematics itself. For this the processes have to be integrated into a coherent whole, essentially that of the standard modelling³/problem solving diagram, a version of which is shown in Figure 1.

³ Mathematical modelling, a useful term, is not yet in everyday use in school mathematics in the UK. Indeed, "modeling" is often thought of as a rather advanced and sophisticated process, used only by professionals. That is far from the truth. We have all been modelling for a long time.

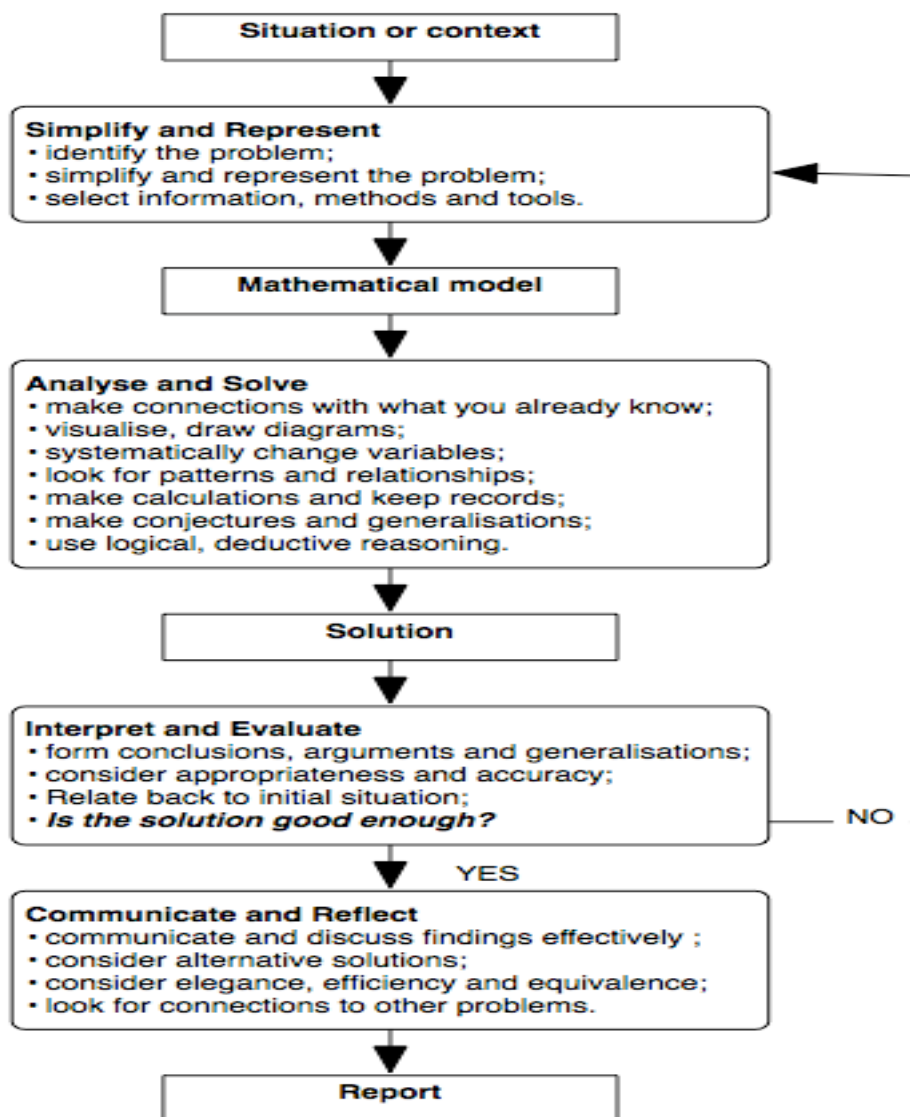


Figure 1. The Modelling Cycle.

Such dangerous ambiguities will be sharply reduced by exemplifying the tasks to be used in assessment, along with their scoring schemes, together with the balance of different task types in the tests (Role B above). The principles and methods for the design of assessment outlined below, if adopted, will ensure that the tasks and tests provide essential support for the classroom implementation of the intentions of the Programme of Study (Role C above). Assessment that covers the range of goals *in a balanced way* will encourage teachers and schools to take these goals seriously, and will reward their students' achievements. The indicator of success, reflecting C, is that *teachers who teach to the test are led to deliver a rich and balanced curriculum*

The argument can be taken further. Do we therefore assess: extended project work; collaborative tasks; practical tasks; oral tasks; computer-based tasks? All these are worth consideration; however, in this paper we focus mainly on what can be, and has been, achieved within the real constraints of timed written high-stakes examinations.

4. Task design principles

The challenge

The classical approach to assessment design is to begin with a matrix listing the various elements of content and the processes to be assessed. Items are then designed to assess each of these elements. This ignores a vital empirical fact – the difficulty of a specific element of performance varies greatly with the task in which it is embedded. Just as scoring goals in shooting practice is easier than in a game of football so, in a complex task involving extended reasoning, students may fail with an arithmetic calculation they could do correctly as a separate exercise. The US Mathematical Sciences Education Board (MSEB 1993), talking of fragmentation, noted:

.... these designs also were often the root cause of the decontextualising of the assessments. If 35% of the items were to be from the area of measurement and 40% of those were to assess students' procedural knowledge, then 14% of the items would measure procedural knowledge in the content domain of measurement. These items were designed to suit one cell of the matrix, without adequate consideration to the context and connections to other parts of mathematics.

In National Curriculum assessments in England since 1989, this fragmentation has been taken to extremes. Consider the following task from the age 16 examination (GCSE):

A triangle has angles $2x$, $3x$ and $4x$.

- (a) Write an expression in terms of x for the sum of the angles.
- (b) By forming an equation, find the value of x .

If a 16-year-old student cannot find x without being led through the task by (a) and (b), is this worthwhile performance in mathematics? Will this fragmentary skill equip the student for subsequent work, or for life? For the student who can do the task without the aid of (a) and (b), this already-simple problem is further trivialized by fragmentation.

This fragmentation arose as the inevitable, but unrecognized, consequence, of adopting the model that separate elements of performance have levels that reflect their difficulty, as specified National Curriculum. Since, in fact, the difficulty of a complex task is not simply that of its parts, the model could only be sustained by testing the parts separately. This is a travesty of performance in mathematics. (If English were assessed in an equivalent way, it would test only spelling and grammar through short items, with no essays or other substantial writing)

It is now accepted that the difficulty of a task depends on the *total cognitive load*, which is determined by the interaction of various aspects of the task including its:

- Complexity
- Unfamiliarity
- Technical demand

as reflected in the chains of reasoning needed for developing a solution. Assessing the level of a student performance on a task has to take all these into account. (This is actually a return to earlier practice in assessing mathematics.)

What does this mean for task design? Let us start by looking at the length of the expected chain of independent reasoning. Compare the triangle task above to the

following task for students in the same age range (see Shell Centre/Joint Matriculation Board 1984, Balanced Assessment 1997–1999 for two versions):

Consecutive Sums

Some numbers equal the sum of consecutive natural numbers:

$$5 = 2 + 3$$

$$9 = 4 + 5$$

$$= 2 + 3 + 4$$

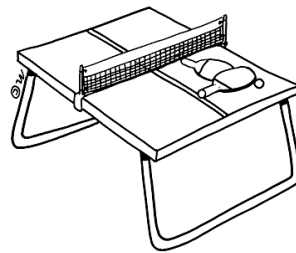
Find all you can about the properties of such “Consecutive Sums”

This is an *open investigation* of a (surprisingly rich) pure mathematical microworld, where students have to formulate questions as well as answer them. Diverse (and incomplete) solutions are expected, and can be used and assessed at various grade levels. It is rather like an essay task on a subject that the student knows about but has not analysed before, let alone been taught “the answer” in class. These are characteristics of thinking with mathematics about problems outside school. The problems there will be more practical but involve the same key processes – problems like *Planning and organising: A table tennis tournament*.

Planning and organising: A table tennis tournament

You have the job of organising a table tennis league.

- 7 players will take part
- All matches are singles.
- Every player has to play each of the other players once.
- There are four tables at the club.
- Games will take up to half an hour.
- The first match will start at 1.00pm.



Plan how to organise the league, so that the tournament will take the shortest possible time. Put all the information on a poster so that the players can easily understand what to do.

Are such problems inevitably more difficult than short items that, in principle, only demand recall? Empirically, the answer is no, or rather, not if they are well-engineered. How can this be? There are a number of factors. When the strategic demand, working out what to try and what mathematical tools to use, is high, the initial technical demand must be lower – in *Consecutive Sums* a lot can be done with simple arithmetic, though some algebra and/or geometry are needed for a fuller analysis. Secondly, it is often easier to tackle a complete problem than parts that have less meaning (Thurston 1999).

Finally, it is clear to the student that *thinking*, connecting with their whole knowledge base, is needed – a very different mode from the usual *attempted recall*. These are not unnecessary complications, as some in mathematics assessment might regard them, but central elements in the performance goals.

The difficulty of a given task can only be reliably determined by trialling with appropriately prepared students – the usual way well-engineered products are developed. If needed, in the light of the designer’s insight and feedback from trials, scaffolding can be added to give students easier access, and a well-engineered ramp of difficulty. We discuss such design tactics in Section 5 below.

Design principles for authentic assessment

Experience across a range of types of assessment suggest the following principles. We have seen above that these principles are often neglected in current UK assessment of Mathematics – indeed, some believe that this is inevitable. However, they are commonplace in other subjects, from which we have much to learn. More directly, there are plenty of examples from past UK public examinations in Mathematics, and those in other countries that are based on similar principles.

1. Assess the types of performance that you want students to be able to do *NOT the separate elements of such performances.* The former also assess the latter, but not vice versa. We shall look at six types of task:

- A. planning and organising
- B. designing and making
- C. modelling and explaining
- D. exploring and discovering relationships
- E. interpreting and translating
- F. evaluating and improving.

Such tasks naturally involve the Key Concepts and Key Processes of the National Curriculum. They assess technical skills at the same time; indeed, the longer chains of reasoning require a higher level of reliability in such skills. To meet the performance goals, tasks have to be *non-routine*, so students know that *thinking* not just *remembering* is needed. None of this means that they have to be more difficult.

It is sometimes argued that substantial tasks of these kinds can be used in the classroom, but not in timed written examinations. This is simply not true; while time constraints impose some limits, examination boards have shown that rich tasks in the 10-20 minute range can assess problem solving, integrating processes and technical skills, in a way that is fair to students (Joint Matriculation Board 1984-88).

2. Assess ‘content’ together with problem solving ‘processes’.

Solving substantial problems requires students to:

- *Represent* problem situations mathematically, stating assumptions
- *Analyse* their model, selecting appropriate mathematical techniques to apply
- *Interpret* and *evaluate* the results
- *Communicate* them in a form suitable for a specified audience

Some tasks will involve all of these processes, some will be more restricted but in a realistic way – for example, interpreting given data and making recommendations, or critiquing a given analysis of a problem.

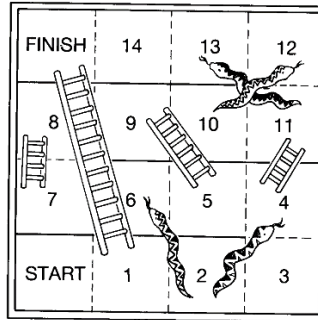
In practice, you cannot assess process without mathematical content – for example, to represent, you have to be able to use representations; to analyse, you need to be able to select and use the appropriate mathematical tools. Further, process and content interact in determining difficulty, which depends on the *total cognitive load* of the task.

Snakes and Ladders

This is a game for two players. You will need a coin and two counters.

Rules

- Take it in turns to toss the coin.
 - If it is heads, move your counter 2 places forward.
 - If it is tails, move your counter 1 place forward.
- If you reach the foot of a ladder, you must go up it.
- If you reach the head of a snake, you must go down it.
- The winner is the first player to reach 'FINISH'.



Suppose you start by tossing a head, then a tail, then a head. Where is your counter now?

List and describe all the faults you notice with the board.

3. Reward reasoning as well as results.

Tasks and scoring schemes should encourage and reward reasoning and explanation, not just correct, accurate answers. As Piaget noted, explanation of a chain of reasoning is an important measure of understanding – and an essential component of mathematics that is functional, where you often need to explain and justify your conclusions. Above we have seen examples of problem types A and D. *Snakes and Ladders*, which has aspects of both B and F, is a task focused on reasoning.

To realise this principle you need scoring schemes that partition the total points among the elements of performance that the task demands, weighted according to their significance in the task. (This is different from using 'method' and 'accuracy' points, an approach common in the UK)

4. Determine the level of a task response by trialling

Trialling is essential for the development of non-routine tasks that work well. The difficulty of a complex task is not simply that of its constituents when tested separately. Difficulty depends in an unpredictable way on a combination of the complexity and familiarity of the task as well as its technical demands, and on the scale of the response needed. Feedback from trialling allows the task design to be refined so as to work well for students across a range of levels.

5. Real world validity

In order to convince students that school mathematics is relevant to their world (an important motivation for many) some tasks should have face validity with students, presenting problem situations:

- that might be of concern or interest to any intelligent and interested citizen;
- in which mathematics can play a significant role in giving better understanding, informing better decisions.

This, again, effectively precludes tasks that have been broken down into steps; life outside the classroom doesn't do that for you. While students have learned to be tolerant of artificial problems in Mathematics, often with contrived "cosmetic realism", these undermine the credibility of the subject. As a result, many people of all ages see secondary school mathematics as irrelevant to their lives; most never use it after they leave school. The contrast with English is stark.

Though many of the tasks will present practical problems, pure mathematical investigations and problems set in fantasy 'other worlds' are also useful, as long as the same processes, using the same mathematical "toolkit", are being assessed. Students enjoy tasks that have an element of drama or surprise, often from unfamiliar associations within the task.

6. Ensure "learning value"

Students who tackle problems that satisfy the 5 principles above will learn something of value from doing so – building new insights and connections in mathematics and/or in the problem context. Since high-stakes assessment is in frequent, why does this matter?

This principle is important for the alignment of assessment and curriculum, establishing the symbiotic relationship between learning and performance. The task types that teachers use in classroom, whether for teaching or assessment for learning, will reflect those on the examination. This principle is also a useful measure of the face validity of the task

These principles have much in common with the way examination tasks used to be designed. However, when the National Curriculum was introduced in the UK, this approach was cast aside for a naïve form of criterion referencing, based on a checklist of "statements of attainment". This precludes using rich tasks, along with many of the other principles set out above. Those constraints will have to be relaxed, if real mathematical performance, including its process aspects, is to be assessed.

However, in assessment design as in all highly skilled activities, it is one thing to describe the principles, it is another to develop the necessary skills to realise them at a high level. The detailed 'engineering' of good assessment tasks is among the more challenging kinds of design and development in education. Hence the following:

7. Find and use a range of exceptional designers

Assessment design of the kind needed to reflect the performance goals is a particularly demanding area of educational design, requiring a wider range of skills and experience than, for example, the design of teaching materials – assessment tasks have to work well without teacher support. To be done well, assessment design needs creative ability. The difference between competent designers of traditional mathematics assessment and that achieved by the best is qualitative. Further, the best designers have characteristic flavours; students deserve the range and variety that can only be achieved by combining the work of several designers.

5. The process of task design: issues, strategies and tactics

Realising these principles in practice raises a range of issues for designers. Here we shall address three important ones: *scaffolding and transparency*, *differentiation*, and *approaches to scoring*. In doing so we shall say something of the strategies and tactics that skilled designers of broad-spectrum assessment of mathematics have developed over the last few decades so as to ensure that their products are well-engineered, i.e. that they work well in realising the intentions in the hands of typical users – in this case, examination providers, examiners, students and their teachers

"Scaffolding" and "transparency".

On the face of it, the assessment of problem solving is straightforward. We set a problem, then assess how well a student can solve it. Difficulties arise, however, when we try to pose the problem in a form which is clear and accessible to all the students, and which elicits useful information regarding their capacity to represent, analyse, interpret and evaluate, and communicate.

A major design issue concerns the level of 'scaffolding' within a task, that is the degree to which students are led through the task, step-by-step. Task designers nearly always have a model solution in mind. They then have to decide how far to guide the student along their solution path. If they do this in a step-by-step fashion, then clearly the task

cannot assess problem solving strategies. If they leave the task 'open' to a wide variety of responses, then students are less clear as to the expectations of the assessor.

Ultimately, we hope that most well-prepared students are able to tackle relatively unstructured problems. However, the design and scoring of open tasks is fraught with difficulty. It is difficult to reconcile 'openness' in the task with 'transparency' in the purpose of the task.

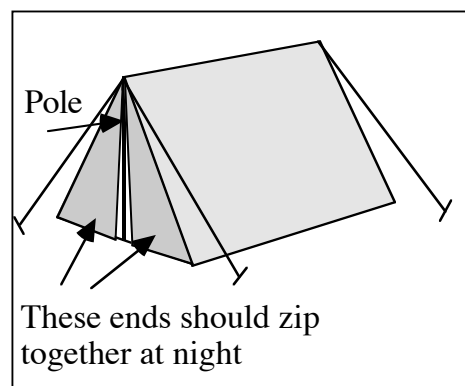
For example, we conceived a task (Figure 2) in which students are faced with the problem of designing a tent, with a triangular cross-section, for two adults to sleep in. The intended response is a drawing showing how the material will be cut, labelled with suitable dimensions.

Figure 2: an open version

Design a Tent

Your task is to design a tent like the one in the picture.

It must be big enough for two adults to sleep in



Draw a diagram to show how you will cut the material to make the tent. Show all the measurements clearly.

In this open format, responses proved difficult to assess, mainly because the task is too ambiguous. Some students design the tent from many pieces of material, while others use a single piece. Some give unrealistic measurements and it is often impossible to say whether this is because they cannot estimate the dimensions of an adult, because they cannot transfer measurements or because they cannot calculate accurately. Some make assumptions that extra space is needed for baggage, but do not explicitly state this. Some use trigonometry, others use Pythagoras' theorem, while others use scale drawing.

With open tasks like this, students often interpret the task in different ways, make different assumptions, and use different mathematical techniques. In fact, they are essentially engaged in *different tasks*. What is more, it is not always possible to infer their interpretations, assumptions and abilities from the written responses. If students have not used Pythagoras' theorem, for example, we cannot tell if it is because they are unable to use it, or simply have chosen not to. This argument may also be applied to mathematical processes. How can we assess whether or not a student can generalise a pattern or validate a solution unless we ask them to?

One solution to the transparency/openness issue, is to clearly define the specific assessment purposes of a *package* of items for the students, making clear what will be

valued – a kind of general rubric⁴. They then know the assessment objectives for the collection, but are not told in a particular task that they should use algebra *now*.

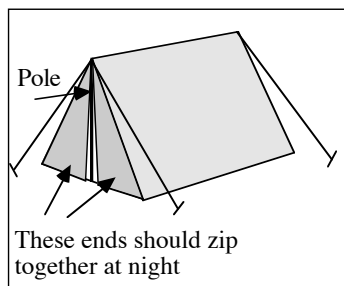
Returning to the tent example, we decided to incorporate more guidance in the task itself. (Figure 3). This version is much more structured. There are clear instructions to estimate, calculate and explain. This enables the assessor to follow through calculations and reasoning. Notice that students are still not explicitly told to use Pythagoras' theorem or trigonometry and a significant degree of problem solving is retained.

Figure 3: a lightly scaffolded version

Design a Tent

Your task is to design a tent like the one in the picture.

Your design must satisfy these conditions:



- It must be big enough for two adults to sleep in (with their baggage).
- It must be big enough for someone to move around in while kneeling down.
- The bottom of the tent will be made from a thick rectangle of plastic.
- The sloping sides and the two ends will be made from a single, large sheet of canvas.
- Two vertical tent poles will hold the whole tent up.

1. Estimate the relevant dimensions of a typical adult and write these down.
2. Estimate the dimensions you will need for the rectangular plastic base. Estimate the length of the vertical tent poles you will need. Explain how you get these measurements.
3. Draw a sketch to show how you will cut the canvas from a single piece. Show all the measurements clearly. Calculate any lengths or angles you don't know. Explain how you figured out these lengths and angles.

To achieve rich and robust assessment, tasks are tried and revised many times, exploring alternative degrees of scaffolding. Another interesting example is given by Shannon and Zawojewski (1995), where trials on two versions of the same task, *Supermarket Carts* are reported. The first, shown in Figure 4, was scaffolded by a series of questions gently ramped in order of difficulty, starting from specific examples to a final, generalised 'challenge'.

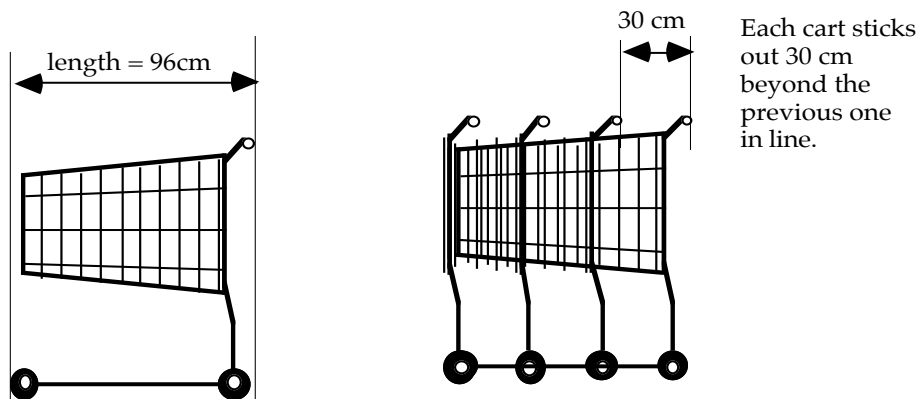
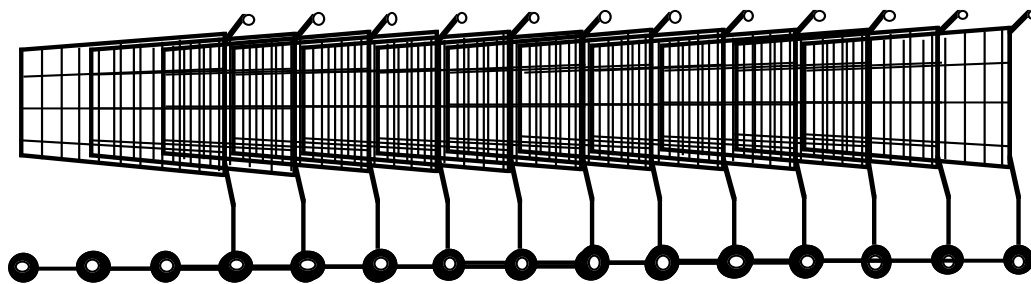
⁴ Initially, these expectations need to be stated explicitly, either in the task or for the package as a whole. Over time, they come to be understood and absorbed

Figure 4.

Supermarket Carts

The diagrams below show 12 supermarket carts that have been "nested" together.

They also show that length of a single supermarket cart is 96 cm and that each cart sticks out 30 cm beyond the previous one in line.



1. How long will a row of 2 carts be? 3 carts? 12 carts?
2. Create a rule that will tell you the length of storage space needed when all you know is the number of supermarket carts to be stored.

Explain **how** you built your rule.

We want to know what data you drew upon and how you used it.

3. Now work out the number of carts that can fit in a space S meters long.

The second version began with a statement of the generalised problem, essentially the last two parts in Figure 4. This study found, as one would expect, that students struggled more with the less structured task and fewer were able to arrive at the general solution. What was perhaps more interesting was that the students perceived the purposes of the tasks as qualitatively different. The students saw the structured task as

assessing content related to equations or functions, while they saw the unstructured task as assessing how they would develop an approach to a problem. Students had no suggestions as to how the structured task could be improved, but they had many suggestions as to how the unstructured could be made to give clearer guidance.

Although they could identify the distinct purposes behind the tasks, they assigned their difficulties to poor task design rather than their own lack of experience of tackling unstructured problems. This result, and many studies of the teaching of modeling skills, emphasize the importance of the alignment between standards, curriculum and assessment, noted earlier – and the associated need for adequately supportive professional development to enable teachers to meet the new challenges involved. The implications of this for implementation are sketched in Section 8

“Differentiation”

An examination must allow all the students who take it to (in Cockcroft’s immortal phrase) *show what they know, understand and can do*, without wasting much examination time in ‘failure activity’. Influenced by the technical limitations of many students in mathematics and the dominance of technical demand in mathematics examinations, he decided that this principle required *differentiation by task* and thus *tiers*. This was strongly opposed by other subjects where *differentiation by outcome only* is standard practice – essay questions allow students to respond at their own, very different, levels and scorers to score them appropriately.

It is important to reconsider this issue for broad-spectrum assessment in mathematics of the kind discussed here. More open tasks allow responses at a wider range of levels. There are design strategies that can help too.

The “exponential ramp” is a powerful design technique for assessing a wide range of levels of performance of students. It uses rich substantial tasks that are scaffolded to offer opportunities at different levels, *increasing the challenge in later parts of the task*.

Consecutive Sums

Some numbers equal the sum of consecutive natural numbers:

$$5 = 2 + 3$$

$$9 = 4 + 5$$

$$= 2 + 3 + 4$$

- Find a property of sums of two consecutive natural numbers.
- Find a property of sums of three consecutive natural numbers.
- Find a property of sums of n consecutive natural numbers
- Which numbers are not “consecutive sums”?

In each case, explain why your results are true.

While this is difficult in technical exercises, it can be done in more open rich tasks. This scaffolded version of *Consecutive Sums* gives students easier access, with a well-engineered ramp of difficulty. Nearly every student can solve the first part; the proof in the last part is challenging for most people.

However, there are always losses as well as gains from scaffolding – here it means that students only have to *answer* questions, not to *pose* them – the latter is an important part of thinking with mathematics.

This and other ways of achieving *differentiation by outcome* on the same task set can obviate or reduce the need for 'tiers', where different students are given more or less challenging tasks according to their perceived level of performance. In the US, where tiers are unacceptable for valid social reasons – essentially the evidence that potentially high-achieving students from less-advantaged backgrounds are put in lower sets with lower expectations. This is not given the same priority in the UK, where it is argued that students should be given tasks that enable them to show what they can do – not what they cannot. An important and interesting dilemma.

Approaches to scoring

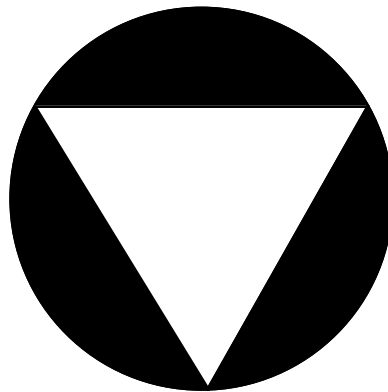
All assessment involves value judgments. The choice of task types defines the range of performances that are valued. Scoring schemes define how far the various elements of performance on a task are valued. Thus scoring, aggregating points and reporting of achievement are major issues in assessment design. Here we shall look at scoring from a broader perspective than is common in UK mathematics assessment.

First we note that the value system is often distorted by the perceived constraints of practicality. Scoring schemes, instead of apportioning credit according to the importance of the elements of performance in the task, assign points to elements that are easy to identify – answers rather than explanations⁵, for example. Tasks are chosen because they are "easy to score", eliminated if scoring may involve judgment. While any high-stakes assessment system must work smoothly in practice, experience in other subjects suggests that many of the constraints that are accepted for Mathematics are unnecessary. We discuss these further in Section 7.

Figure 5

Magazine Cover

This pattern is to appear on the front cover of the school magazine.



You need to call the magazine editor and describe the pattern as clearly as possible in words so that she can draw it.

Write down what you will say on the phone.

⁵ Some GCSE scoring schemes give full points for correct answers, even when working or explanation is asked for – the latter is merely evidence for partial credit if the answers are wrong. This ignores the value of explanation as evidence of mathematical understanding.

Figure 5 shows Magazine Cover – designed for ages 8-10, it challenges older students. The scoring scheme uses a fairly standard point system. The total points available are chosen to be equal to the length of time (in minutes) it takes a typical successful student to complete the task. This arbitrary choice reflects the need for precision without overloading examiners’ judgments with too much detail. The total points for each task are then distributed among the different aspects of performance, so that each aspect is given a weight appropriate to its importance.

This process is illustrated in the rubric design for *Magazine Cover*, which shows the various elements that are credited.

Magazine Cover		Points
<div style="border: 1px solid black; padding: 5px; display: inline-block;"> The core elements of performance required by this task are: <ul style="list-style-type: none"> • describe a given geometric pattern </div>		
Based on these, credit for specific aspects of performance should be assigned as follows		
A circle.		1
A triangle.		1
All corners of triangle on (circumference of) circle.		1
Triangle is equilateral. Accept: All sides are equal/the same.		1
Triangle is standing on one corner. Accept: Upside/going down.		1
Describes measurements of circle/triangle.		1
Describes color: black/white.		1
<i>Allow 1 point for each feature correctly described up to a maximum of 6 points.</i>		
Total Points		6

The maximum score is 6 because this is a 6- minute task. For each of the two simple technical terms, circle and triangle, 1 point is assigned. (Most students get these points.) For each of the key geometric insights on the triangle (equal sides, point down, touching the circle), 1 point is given, but further technical terms (e.g., equilateral) are not required at grade 3. For the color contrast, seen as a key feature in the context of the task (cover design), 1 point is given, and 1 point is given for some size information. To keep within the total, a maximum of 6 of the 7 points above may be awarded.

In more complex tasks, more than one point may be assigned for an important and substantial part of the task, such as for an explanation. The rubric will give guidance on assigning partial credit for a correct but incomplete explanation.

There are value judgments throughout this and every other rubric. Here, for example, some mathematics teachers criticise the point for colour – “not mathematics”, but important in the context of the problem. Others want some credit for the clarity of the explanation⁶, or note that the editor will determine the size. The Programmes of Study give some guidance on values but more discussion and decisions will surely be needed.

⁶ German mathematics teachers demand that explanations are in good German

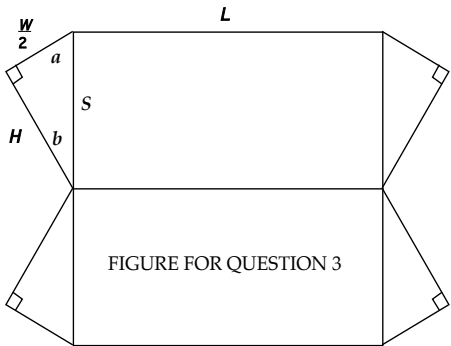
Alternative approaches to scoring are worth considering. Here we will explore two aspects:

- Profile scoring, under categories
- Holistic v point scoring

There are advantages and disadvantages in each of these options. We will exemplify them together

Pointwise scoring involves allocating points to particular attributes of a response, then combining these to obtain a single score or multiple scores within different categories. Usually, the process of devising such a scheme begins with allocating points to each step of a model solution, then adjusting for alternative responses, 'following-through' errors, and so on. In use, the advantages of this method are that each judgment is small, so the scoring task can be easily absorbed and implemented quickly and reliably by relatively inexperienced scorers. Its major disadvantage is that numerical scores do not communicate the essential quality of the response.

Figure 6: Design a Tent: a 'pointwise by category' scoring scheme.

	Correct response	PSR	N&Q	G
Qu1	Suitable height, breadth, kneeling height of person estimated.		3	
Qu2	Follow through students answers from from Qu 1. Length (L) of base = height of person + k ($k > 0$) Width (W) of base = 2 x breadth of person + c ($c > 0$) Kneeling height \leq Height (H) of tent poles \leq Height of person.			1 1 1
Qu3	Follow through students answers from from Qu 2. <ul style="list-style-type: none"> • Correct number of tent faces shown • Faces joined together correctly • Right angle formed by vertical zip and ground shown correctly • Length L consistent with base (Qu2) • End flaps consistent with base ($W/2$ and H) • Appropriate method <i>selected</i> to calculate S • Method applied to calculate S correctly • Correct answer obtained for S. • Appropriate method <i>selected</i> to calculate a or angle b • Method applied to calculate a or b correctly • Correct answer is obtained for angle a or b • Correct answer obtained for second angle  <p style="text-align: center;">FIGURE FOR QUESTION 3</p>	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1 1 1	
Totals		12	3	15

Profile scoring There are some advantages in scoring and reporting different aspects of student performance separately – Concepts, Skills and Problem solving form a popular trio in the US. This allows teachers to see areas of strength and weakness of individual students, and of their own teaching. It allows school systems to look at progress more deeply, particularly as new goals (like problem solving) are introduced into the system. Data on the effect of the new broader US teaching materials on performance show roughly the same standards in mathematical skills but real gains in problems solving.

Let us look at the comparison for the *Design a Tent* task. Figure 6 contains a sample "pointwise by category" scoring scheme (Balanced Assessment 1995). This assigns points to separate categories, which are aggregated separately to give a *student profile*.

Here features of the response are credited under three categories; Problem solving and reasoning (PSR): 12; Number and Quantity (N&Q): 3; Geometry (G): 15. On some occasions, where there is a big strategic demand within the task, points will be given for both content and process categories. This does not amount to double weighting as the points in different categories are not added together but reported separately as a profile.

Holistic scoring involves considering each response as a single entity and comparing its quality with 'benchpoint' descriptors and sample responses. Such a scheme may be devised empirically, by categorising actual responses, then by writing descriptions which aim to capture the essential features that are common to categories. Figure 7 offers a sample holistic scheme for the same task and Figure 8 gives a sample piece of work pointed according to both schemes.

Figure 7: Design a Tent: a task-specific holistic scoring scheme

1. *The student needs significant instruction*

Typically the student understands the prompt and attempts to estimate dimensions and draw diagrams. The student is unable to produce satisfactory estimates or coordinate the constraints in the problem. The student may attempt to transfer measurements to a drawing of base of the tent, but is unable to visualise how the top may be constructed.

2. *The student needs some instruction*

Typically the student attempts to satisfy some but not all of the constraints in the problem. Some reasonable estimates are made. The student attempts to show how the tent may be constructed and transfers some measurements correctly to a drawing. There is no attempt to calculate new lengths or angles.

3. *The student's work needs to be revised*

Typically, the student attempts to satisfy all constraints in the problem. The student attempts to show how the tent may be constructed and transfers measurements correctly to a drawing. The student selects and uses appropriate mathematical techniques to calculate new lengths or angles.

A suitable tent could not yet be successfully constructed from the plan.

4. *The student's work meets the essential demands of the task.*


Typically, the student satisfies all constraints in the problem. The student shows how the tent may be constructed and transfers all measurements correctly to a drawing. Appropriate mathematical techniques have been used to calculate new lengths and angles. These calculations are mostly correct.

A suitable tent could be constructed from the plan.

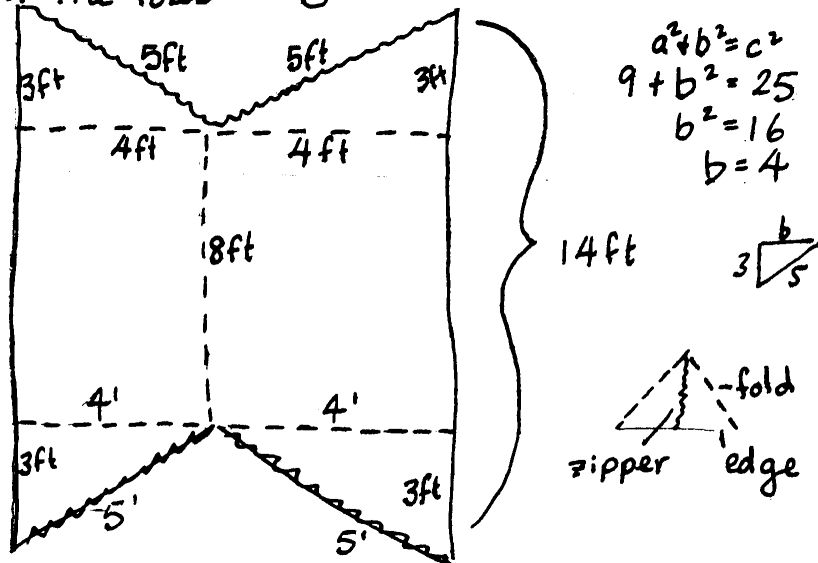
The advantage of this method is that the score assigned is focused on the overall quality of the response, which is not always captured by point scoring. The disadvantage

is that the scorer has to internalise more information and the judgments are broader and more subjective. Though holistic scoring has been used for high-stakes assessment in Mathematics (New Standards 1995), and is common in other subject areas, its main strength in the UK Mathematics is probably for formative assessment, where a generic (i.e. not task specific) scoring scheme can cover any tasks teachers or students choose.


Figure 8: Sample student's work

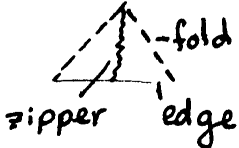
Design a Tent 

1. Size typical adult ^{TALL} 6ft x ^{WIDE} 2ft
2. 6ft wide & 8ft long
2 extra feet in the length & 2 extra feet in width more than average person two people to give room in between the people
- 3 Length of the Poles 5'



$a^2 + b^2 = c^2$
 $9 + b^2 = 25$
 $b^2 = 16$
 $b = 4$

14ft 



This student has correctly estimated the dimensions of a typical adult and designed an appropriate plastic base for the tent. Her linear dimensions are transferred correctly to the plan but she has not identified the right angle at the base of the zip correctly. She realizes that the Pythagorean theorem is appropriate. No attempt has been made to calculate angles.

On the **pointwise** scheme, she scores as follows:

Problem solving and Reasoning:	5/12
Number and Quantity	2/3
Geometry	7/15

On the **holistic** scheme, she is near the borderline at the top of category 2. While she has made an error in transferring the right angle to her diagram (typical of category 2), she has shown that she can select and use Pythagoras' theorem, but not trigonometry (typical of category 3).

We find both types of scoring informative and useful. It is possible to obtain an approximate correspondence between the two schemes and scoring a student response in both ways gives interesting insights. It is also possible to make students aware of the holistic scheme while they are doing the task, thus enabling them to see its purpose more clearly.

6. Building tests from tasks

The assembly of tasks of different types and lengths into tests is an often-contentious process that brings out differences in value systems. The following heuristic strategies have been found to yield high quality tests. We pick them out because current mathematics tests clearly do not use them. Their essence is a two-step process, separating the creative process of task design and development from the analytical process of balancing tests.

- 1 Give the designers an open brief to design good tasks.** Fine designers have internalised the essential elements of task design, and the domain for which they are assessing. We have found that the most effective, and most admired, tasks emerge when the designer is given only broad guidance – the range of process and content to be covered in the test, the students involved, the total time,... – rather than a detailed brief for each task. (That more common approach⁷ is usually designed to sample the *analytical domain description* ‘matrix’ rather than *performance in the subject*.)

The resulting collection of tasks can then feed into a selection and test assembly process. Using several sources of tasks, designed to a common brief, can give variety of challenge

- 2 Analyse tasks against the domain framework.**

Constructing tests that are balanced across the various dimensions of the accepted framework in accordance with the overall learning goals and constraints of time and circumstances of performance is fundamentally an analytical challenge. MARS has developed a *Framework for Balance* (Figure 9 overleaf) that structures this process. It makes explicit the weights of the different aspects of performance that each task demands: processes, content areas, task length, openness and non-routine aspects. It thus supports a process of balancing tests across these various dimensions (often, otherwise, only content areas are balanced).

Adapted to the Programmes of Study, this approach would help discussions and decision within the test design group and, more importantly, help counter attacks on an innovative test through the inevitable, detailed criticism of the exemplar tests.

⁷ In the design of Key Stage Tests and GCSE in Mathematics, designers face specific constraints. “Design tasks with 20% of the points on each Attainment Target, and with 30% at level 4, 40% at level 5, 30% at level 6”. Both the products and experience show that such constraints inhibit designers and lead inevitably to poorer tasks, notably short items that are a travesty of performance in mathematics.

Figure 9. A Framework for Balance
(MARS 1998, adapted for QCA Programme of Study)

Mathematical Content Dimension

- *Mathematical content* will include some of:

Number and quantity including: concepts and representation; computation; estimation and measurement; number theory and general number properties.

Algebra, patterns and function including: patterns and generalization; functional relationships (including ratio and proportion); graphical and tabular representation; symbolic representation; forming and solving relationships.

Geometry, shape, and space including: shape, properties of shapes, relationships; spatial representation, visualization and construction; location and movement; transformation and symmetry; trigonometry.

Handling data, statistics, and probability including: collecting, representing, interpreting data; probability models – experimental and theoretical; simulation.

Mathematical Process Dimension

- *Phases* of problem solving, reasoning and communication will include:

Representing (Modelling and formulating)

Analysing (Transforming and manipulating)

Interpreting (Inferring and drawing conclusions)

Evaluating (Checking and evaluating)

Communicating (Reporting)

Task Type Dimensions

- *Task type*: open investigation; non-routine problem; design; plan; evaluation and recommendation; review and critique; re-presentation of information; technical exercise; definition of concepts.
- *Non-routineness*: context; mathematical aspects or results; mathematical connections.
- *Openness*: open end with open questions; open middle.
- *Type of goal*: pure mathematics; illustrative application of the mathematics; applied power over the practical situation.
- *Reasoning length*: expected time for the longest section of the task. (An indication of the amount of scaffolding),

Circumstances of Performance Dimensions

- *Task length*: short tasks (5-15 minutes), long tasks (15-60 minutes), extended tasks (several days to several weeks)
- *Modes of presentation*: written; oral; video; computer.
- *Modes of working*: individual; group; mixed.
- *Modes of response*: written; built; spoken; programmed; performed.

7. Constraints on assessment design: myths and realities

Examination boards are properly concerned with the practicalities of their tests – and so are we. They have a long list of reasons why desirable things cannot be done. Some of these constraints are unavoidable but extensive experience over many decades shows that others are not. In this section we address some of these concerns, distinguishing the inevitable from the unnecessary.

Time

Myth 1: Testing takes too much time.

Feedback is a vitally important factor in improving performance – in sport, “games” give purpose to training, the purpose of music practice is to play “gigs” (or concerts for the less ‘cool’). If doing the assessment tasks is also good learning, that is a further justification for spending time on assessment.

We accept that, at least in the current UK and US political climate, constraints on the time for high-stakes assessment are inevitable⁸. The issues include:

- How much time is reasonable?
- How can it best be used?

On the first, practice around the world ranges widely – at least from about 40 minutes to many hours per year. A reasoned decision would relate this to both the time for structured classroom assessment and for teaching. The pressures for reducing the time reflect the distaste for substantial testing that many key groups share – though for very different reasons.

Sampling

Myth 2: Each test should cover all the important mathematics.

In mathematics people say: “We taught (or learned) X but it wasn’t on the test.”

Mathematics is the only subject where there is a tradition of “coverage”, assuming that all aspects of new content should be assessed every time. This has been at the expense of any significant assessment of process aspects; once the interaction between process and content in tasks is recognized, it is clearly impossible to assess the full range of types of performance. Is this a concern?

Sampling is accepted as the inevitable norm in all other subjects. History examinations, year-by-year, ask for essays on different aspects of the history curriculum; the same is true in, say, chemistry; final examinations in literature or poetry courses do not expect students to write about every set book or poem studied. It is accepted that a given examination should:

- Sample the domain of knowledge and performance
- Vary the sample from year to year, so that teaching addresses the whole domain
- Emphasise aspects that are of general importance, notably the process aspects

However, the balance of the sampling is crucial, discussed in Sections 2 and 6, answering

Myth 3: “We don’t test that but, of course, all good teachers teach it.”

⁸ Comprehensive evaluation in the 1980s of the “100% coursework” assessment schemes for English, with no timed tests, showed that they satisfied all reasonable requirements of reliability and fairness; however, the then Prime Minister’s decision to outlaw them reflected a widespread “common sense” perception to the contrary. Mathematics coursework has suffered from inadequate engineering.

Accuracy

Myth 4: Tests are precision instruments.

They are not, as test-producers' fine print sometimes makes clear. Mathematics examiners have long been proud of the consistency between different pointers – the typical *point-repoint variation* is often just a point or two, much less than in some other subjects. However, that is not a fair measure of accuracy of what a student knows, understand and can do; that must include the *test-retest variation*. Testing and then retesting the same student on parallel forms, "equated" to the same standard, should produce the same scores. In fact, they are likely to be substantially different. There is a reluctance to publicise test-retest variation, or even to measure it⁹. Most estimates suggest that the total uncertainty is roughly one grade – on retest, a Grade C is as likely to be either a B or a D as a C. This fact is ignored by policy makers who know that measurement uncertainty is not politically palatable, when life-changing decisions are made on the basis of test scores.

Tests are fair, but in the same way as a lottery is fair.

What are the implications of this for assessment design in Mathematics? The drive for "precision" has led to narrow *de facto* assessment objectives and simplistic tests. This is clearly pointless – the true uncertainties remain high and the price paid from unbalanced assessment is as unnecessary as it is harmful. Mathematics should be content with point-repoint variation comparable with other subjects, notably English, which command public confidence and respect. With the kinds of task described in this paper, that is readily achieved.

8. Implementing assessment improvements

If the improvements in assessment along the lines outlined in this paper are to become a reality, how can this be achieved? There are so many examples of gross mismatch between the outcomes and the intentions of sensible policy decisions that it is vital to recognize the scale of the challenge. How best to do so is a huge subject; here we shall content ourselves with raising some issues and offering comments suggestions.

This development challenge is non-routine Some policy changes lie within the competence and expertise of those most affected; these can safely be designed by practitioners and implemented after piloting. Others lie outside the range of current practice; these need a different approach – a research-based development effort by teams with a proven track record of related innovation. In education, practitioners rarely acknowledge their limitations – neither teachers leaders nor examination boards like to say "We don't know how to do this" (Other professions, doctors for example, are more realistic.) Work in England so far on "Functional Mathematics" assessment, which needs similar treatment to that outlined here, exemplifies this problem. Government regularly funds innovative projects in non-mainstream areas (e.g. World Class Tests for gifted students); it is even more important to do so in crucial areas. The national tests for ages, 7, 11 and 14 have been systematically developed though, in Mathematics, with too narrow an apparent brief; in contrast, the more important GCSE examinations in Mathematics have methods of design and development that, while effective for replicating minor variations on earlier tests, are unsuited for innovative challenges.

There is much more to say about the methodological implications of tackling innovative design challenges. Here we will only say that systematic research-based design and

⁹ One rigorous recent study (Gardner 2002??), looked at an "eleven plus" test in Northern Ireland, a traditional test with important "consequences". Testing the same cohort with two equivalent tests, they found that, of half the 6000 students who "passed" on one test, about 3000 would be different if the other test were used. Pressure was placed on the authors, and their university, not to publish the results

development by several teams working in parallel to a common brief is the approach most likely to yield high-quality outcomes¹⁰.

Pace of change: "big bang" v incremental When a problem is identified, there is a political urge to solve it. When the solution implies profound changes, particularly in the well-grooved practice of professionals, the rate at which they can change without corruption of the intentions is an important factor. The level of support, of proven effectiveness, is also important. If the pace of change required is too great (an empirical question), important aspects of the planned change may simply "disappear into the sand". This has been a recurrent problem with "big bang" implementations. An alternative model (Joint Matriculation Board/Shell Centre 1982-86), based on regular well-supported incremental changes, has proven effective in implementing profound changes over a few year period using aligned materials for assessment, teaching and professional development. This proved effective and popular with teachers – a significant factor in faithful implementation. It also avoids any public upheaval in tried and trusted examinations – a crucial advantage. In the current context, this implies introducing new task types gradually over a period of a few years until the target balance is achieved.

The challenge to the examination boards The English exam boards have both public service and commercial roles, with an expectation that they will deliver their examinations and results impeccably. While the former makes them anxious to "do the right thing", they are likely to lose market share if their tests are perceived as more difficult – a likely effect of any profound change. Thus they all have a commercial incentive to minimise any such change.

What might be done to avoid distortion of the outcomes that policy seeks? A number of possibilities are worth considering. The design of the new types of task might be assigned to other agencies. The selection of a mandated proportion of tasks from this "bank" could be left to individual boards, or could be given to another national body.

It could be decided that all providers include in their papers the same set of novel tasks. Common tasks of this kind would remove any question of a "race to the bottom". It would have the additional advantage of providing some comparative information on the overall standards of the different boards' examination based on a subset of tasks of high validity. (This might allow QCA to do less micromanagement of board examinations.) The common questions approach was considered a decade ago; while no insuperable objections emerged, it was unpopular with the exam boards, who are proud of their independence and omnicompetence.

References and appendices

To come

¹⁰ The Bowland Trust, with DCSF support, has taken this kind of approach to the development of "case studies" on real problem solving– teaching units that closely reflect the Programme of Study, are supported by a linked professional development package. Assessment development is next.