# Assessing Mathematical Proficiency: What is important?

Hugh Burkhardt
MARS: Mathematics Assessment Resource Service
Nottingham and Michigan State Universities

My brief is to look at those aspects of performance in mathematics that are important, and to illustrate how they may be 'measured' in practical assessments of students K-12, both high-stakes and in the classroom. To this end, this paper addresses the following questions:

Who is assessment for? *Students? Teachers? Employers? Universities? Governments?*

What is it for? *To monitor progress? To guide instruction? To aid, or to justify, selection?*

**What mathematics values should assessment reflect in its tasks and scoring?**

When should it happen to achieve these goals? *day-by-day? monthly? yearly? once?*

What will the consequences be – *for students, teachers, schools, parents, politicians*?

What will it cost, and is this an appropriate use of resources?

There are, of course, multiple answers to each of these inter-related questions, requiring a mix of different kinds of assessment: summative tests, assessment embedded in the curriculum, and informal observation and feedback in the classroom, day-by-day. Rather than discussing them one-by-one, we shall concentrate in this paper on principles that should guide the choice of answers, particularly to the third question: *What is important in doing mathematics*? All assessment is based on a system of values, often implicit, where choices have to be made (see, for example, National Research Council 2001); here I seek to unpack these, so the choices can be considered and explicit.

The discussion will mix analysis with illustrative examples. Specific assessment tasks are, perhaps surprisingly, a clear way of showing what is intended – a short item cannot be confused with a long, open investigation, whereas "show a knowledge of natural numbers and their operations" can be assessed by either, representing very different kinds of performance.

## Assessment design principles

**Measure what is important***, not just what is easy to measure* This is a key principle – and one that is widely ignored. Nobody who knows mathematics thinks that short multiple-choice items really represent mathematical performance; rather, many believe it doesn't matter much what kinds of performance are assessed, provided the appropriate mathematical topics are included. The wish for cheap tests that can be scored by machines is then decisive, along with "Math tests have always been like this"[1]. This approach is widely shared in all the key constituencies, but for very different reasons. Administrators want to keep costs down. Psychometricians are much more interested in the statistical properties of items than what is assessed, and the assumptions underlying their procedures are less-obviously flawed for short items. Teachers dislike all tests and want to minimise the time spent on them as a distraction from "real teaching" – ignoring the huge amounts of time they now spend on test-prep that is not useful for learning to do mathematics. Parents think multiple choice tests are "fairer" than human scoring, ignoring the

[1] Only in the US, particularly in mainstream K-12 education. Other countries use much more substantial tasks, reliably scored by people using rigorous scoring schemes. US AP exams and university assessment are also like this.

values and biases that this kind of task introduces. None are aware of the damage this kind of assessment does to students' learning of, view of and attitude to mathematics. This paper is about how one can do better. The second principle follows:

**Assess performances that you are interested in**, *not just the separate ingredients of such performances*. Measuring the latter tells you little about the former – because, in most worthwhile performance, the whole is much more than the sum of the parts. Do they assess basketball only through 'shooting baskets' from various parts of the court and dribbling and blocking exercises? Of course not – they watch the player in a game. Do they assess a pianist only through scales, chords and arpeggios (though all music is made of these)? Of course not – though these may be part of the assessment, the main assessment is on the playing of substantial pieces of music. Mathematical performance is as interesting and complex as these, and should equally be assessed holistically as well as analytically. When we don't (and, in mathematics assessment, that is currently most of the time), is it any surprise that so many students aren't interested; no intelligent music student would choose a course on scales and arpeggios.

**What does this imply for K-12 mathematics?** Consider the following simple task:

*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 by (a) and (b), is that worthwhile mathematics? For the student who can, this already-simple problem is further trivialised by fragmentation. Compare this to the following task (Balanced Assessment 1997-99), for students of the same age:

*CONSECUTIVE ADDENDS*

*Some numbers equal the sum of consecutive natural numbers*

*5 = 2 + 3*

*9 = 4 + 5*

*= 2 + 3 + 4*

- *Find out all you can about consecutive addends*

This is an *open investigation* of a surprisingly rich pure mathematical microworld, where students have to formulate questions as well as answer them. It is a truly *open-ended* [2] task, ie one where diverse (and incomplete) solutions are expected, and can be valued at various levels. Scaffolding can be added to give students easier access, and a well-engineered ramp of difficulty.

- *Find a property of sums of two consecutives*
- *Find a property of sums of three consecutives*
- *Find a property of sums of n consecutives*
- *Which numbers are not consecutive addends?*

*In each case, explain why your results are true*

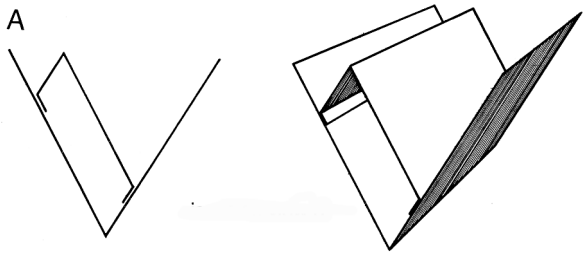
The proof in the last part is challenging for most people. However, the scaffolding means students only have to *answer* questions, not to *pose* them – an essential part of doing mathematics. Is this the kind of task 16-year old students should be able to tackle effectively?

[2] not just *constructed response*, a common confusion and a critical difference

What about the following task (Shell Centre 1987-89)? Is it worthwhile, and worthwhile mathematics?

## *WILL IT FOLD FLAT?*

*Diagram A is a side view of a pop-up card.*

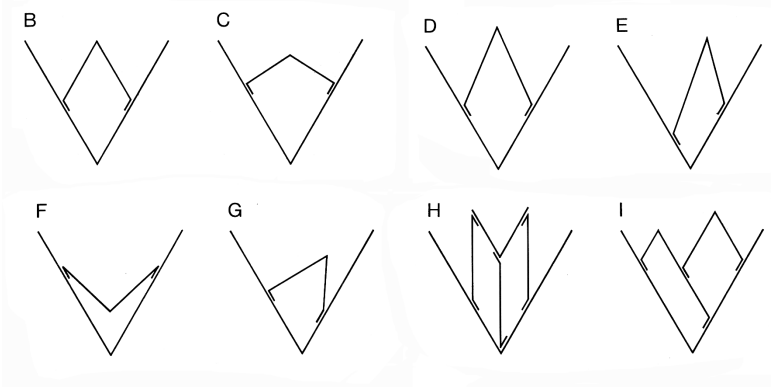



*Look at the diagrams below*

*Which cards can be closed without creasing in the wrong place?*

*Which can be opened flat without tearing?*

*Make up some rules for answering such questions.*

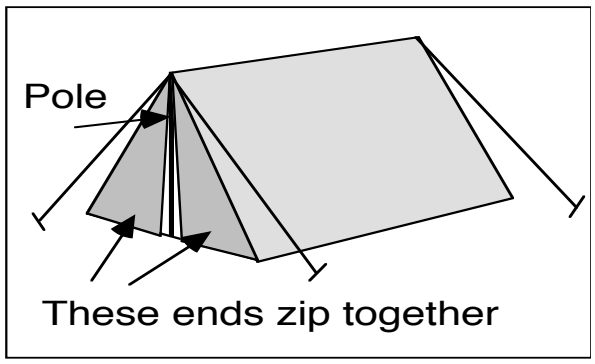



What about the following, more practical, task?

## *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 (with baggage).*

*The tent should be big enough so that someone can move around while kneeling down. Two vertical tent poles will hold the whole tent up.*



Again, would the following more-scaffolded version of the prompt be a more suitable performance goal, or does it "lead them by the hand" too much? (Feedback in development of tasks with students guides such design decisions.)

- *Estimate the relevant dimensions of a typical adult.*
- *Estimate the dimensions of the base of your tent.*
- *Estimate the length of the vertical tent poles you will need.*
- *Show how you can make the top and sides of the tent from a single piece of material. Show all the measurements clearly.*

*Calculate any lengths or angles you don't know.*
*Explain how you figured these out*

This is a typical fairly closed *design task*, requiring sensible estimation of quantities, geometric analysis and numerical calculations (even the Pythagorean Theorem).

These tasks (particularly the last two) are also seen as worthwhile by people who are *not* mathematicians or mathematics teachers. (Most people will not become either – but they *all* have to take high school math) The choice of performance targets, illustrated by the above exemplars, is at the heart of defining K-12 mathematics. All sectors of society have an interest in these choices; mathematicians and mathematics educators need their views, and their informed consent. This implies the kind of well-informed debate that remains rare – and, too-often, is obfuscated by the emotional simplistics of both 'left' and 'right' in the 'math wars'.

**Correlation is not enough** It is often argued that, though 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 depends on what you mean by "correlate well"), it is *not* a justification for narrow tests. Why? Because assessment plays *three* major roles:

- **A: to 'measure' performance** – ie
  *"to enable students to show what they know, understand and can do"*

but also, with high-stakes assessment that impacts students' and teachers' lives, **inevitably**

- **B: to exemplify the performance goals** – assessment tasks communicate vividly to teachers, students and their parents what is valued by society, and thus
- **C: to drive classroom learning activities** *(WYTIWYG: What You Test Is What You Get)*

These roles carry responsibilities for test designers. Correlation is never enough, because it only recognizes A. The effects through C of cheap and simple tests of short multiple-choice items can be seen in every classroom – the fragmentation of mathematics, the absence of substantial chains of reasoning, the emphasis on procedure over assumptions and meaning, the absence of explanation and mathematical discourse,... The list goes on.

**Balanced assessment** accepts all three responsibilities A, B and C. They imply that assessment should be designed to have two properties:

- **Curriculum balance**, such that teachers, who will "teach to the test", are led to provide a rich and balanced curriculum covering *all* the learning and performance goals that the standards (state, national and/or international) embody.
- **Learning value** – because such high-quality assessment takes time, the assessment tasks should be worthwhile learning experiences in themselves.

Assessment with these as prime design goals will, through B and C above, support rather than undermine learning high-quality mathematics. This is well-recognized in some other countries, where assessment is used to actively encourage improvement. In the US, a start has been made – MARS (2000–) and New Standards (1998) have developed better-balanced assessment, as have some states. However, cost considerations too-often lead school systems to choose cheap multiple-choice tests that assess only fragments of mathematical performance – despite the tiny fraction of educational spend that goes on assessment.

It thus follows from B and C that choosing the range of task types that you will assess is also a rather clear way of defining a curriculum. (Lists of math content, while essential, evade most of the key issues about performance: What should be the balance of short items, 15 minute tasks, or three week projects, for example?) More on this, and how it relates to a more analytic approach, later.

**Some common illusions** about assessment are worth noting:

- *Tests are precision instruments;* they are not, as the providers' fine print usually makes clear. Testing, then retesting, a student on parallel forms, 'equated' to the same standard, produces significantly different scores. This is ignored by most test-buyers who know that measurement uncertainty is not politically palatable, when life-changing decisions are taken on the basis of test scores. The drive for precision leads to narrow assessment objectives and simplistic tests. (The logical end is to measure each student's height, which is well-correlated with math performance for students across the age range 5-18)
- *Each test should cover all the important mathematics* It does not and cannot, even when you narrow the range of mathematics to short content-focussed items; it is always a sampling exercise. This does not matter as long as the samples in different tests range across all the goals but some object: "We taught/learnt X but it wasn't tested this time". (Such sampling is accepted as the inevitable norm in other subjects. History examinations year-by-year, will ask for essays on different aspects of the curriculum.)
- "*We don't test that but, of course, all good teachers teach it.*" If so, then there are few "good teachers"; the rest take very seriously the measures by which society chooses to judge them and, for their own and their students' futures, concentrate on these.
- *Testing takes too much time* Feedback is important in every system; below we shall look at the cost-effectiveness of assessment time.

## What should we care about?

We now take a further look at this core question. Is "Will these students be prepared for our traditional undergraduate math courses?" still a sound criterion for judging K-12 curricula and assessment? What other criteria should be considered? (Personal viewpoint: the traditional imitative algebra-calculus route was a fine professional preparation for my career as a theoretical physicist[3]; however, for most people it is not well-matched to their future needs – except for its "gatekeeping" function which could be met in various ways[4]) In seeking a principled approach to goal-setting, it is useful to start with a look at societal goals – what capabilities people want kids to have when they leave school. Interviews with widely differing groups produce

[3] Not surprising, since it was essentially designed by Isaac Newton – and not much changed in content since.

[4] Latin was required for entrance to both Oxford and Cambridge Universities when I was an undergraduate. All now agree that this is an inappropriate gatekeeper.

surprisingly consistent answers, and their priorities are not well-served by the current math curriculum. I have space to discuss just one key aspect.

**Automata or Thinkers?** Which are we trying to develop? Society's demands are changing, and will continue to change, decade by decade – thus students need to develop flexibility and adaptability in using skills and concepts, and in self-propelled learning of new ones. US economic prosperity depends on developing *thinkers* at all levels of technical skill, whether homebuilder, construction-site worker, research scientist or engineer. Equally, it is absurd economics to spend the ~$10,000 cost of a K-12 education in mathematics developing the skills of *automata* that can be bought for ~$5 - $200. *Thinkers* also have more fun than *automata*, which is important for motivation. How do we assess *thinkers*? We give them problems that make them *think*, strategically, tactically and technically – as will many of the problems they will face after they leave K-12 education, where mathematics can help.

**Mathematics – inward- or outward-looking?** Mathematicians and many good mathematics teachers are primarily interested in mathematics itself. For them, its many uses in the world outside mathematics are a spin-off. They are two admirable professions, doing important work -- but they are a tiny minority of the population, in school and in society as a whole. They rightly have great influence on the design of the mathematics curriculum K-12, but should the design priorities be theirs, or more outward-looking ones that reflect society's goals? The large amount of curriculum time devoted to mathematics arose historically from its utility[5] – the value for everyone of mathematics to use in the outside world. That priority, which now implies different mathematics, should continue to be respected.

**Mathematical Literacy** is an increasing focus of attention, internationally (see eg PISA 2003) and in the US (see eg Steen 2000). PISA, the OECD *Programme of International Student Assessment,* seeks to assess mathematical literacy (ML), complementing TIMSS which is mathematically inward-looking. Various terms[6] are used for ML: in the US Quantitative Literacy (QL) is common ; in the UK, where the term *numeracy* was coined for ML (Crowther Report 1959), it is now (Tomlinson Report 2000) being called Functional Mathematics(FM). We describe it thus:

> *Functional mathematics* is mathematics that most *non-specialist adults*,
> if they are taught how, *will benefit from using in their everyday lives* to better understand and operate in the world they live in, and to make better decisions.

Secondary school mathematics is non-functional for most people. (If you doubt this, ask any non-specialist adult, such as a teacher of English Language or an administrator, when they last used some mathematics they first learned in secondary school). ML=QL=FM is distinct from the

> *Specialist Mathematics* that is important in education for various professions.

All the current curriculum is justifiable as specialist mathematics for some professions. However, as a gatekeeper subject, which is a key part of the educational record of everyone, should it not have a large component of functional mathematics that every educated adult will actually use?

---

[5] The argument that mathematics is an important part of human culture is clearly also valid – but does it justify more curriculum time than, say, music? Music currently gives much more satisfaction to more people.

[6] Each of these terms each has an inherent ambiguity. Is it literacy *about* or *using* mathematics? Is it functionality *inside* or *with* mathematics? It is the latter that is the focus of those concerned with mathematical literacy.

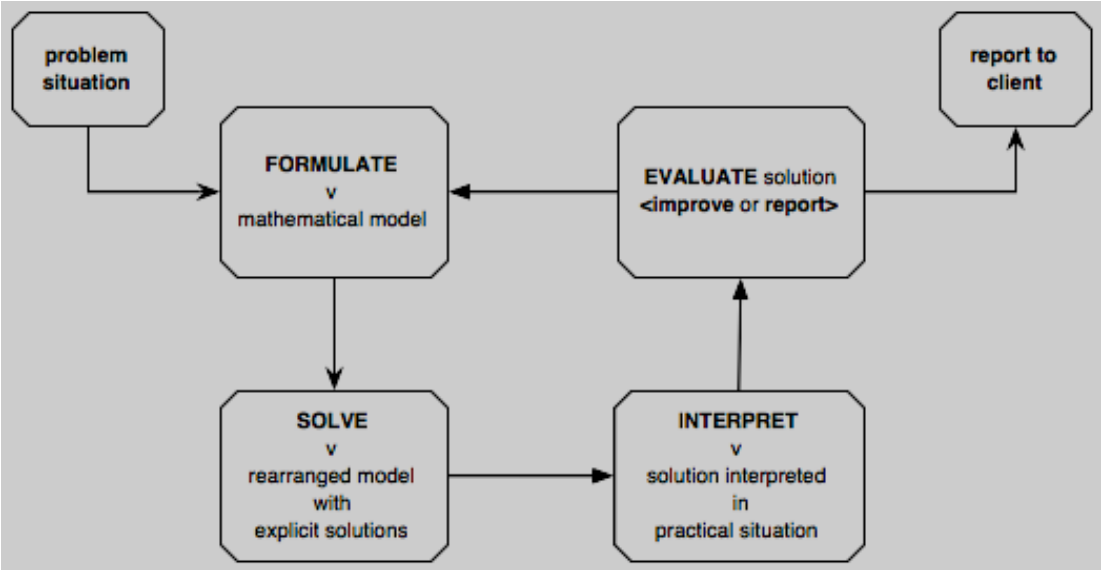
I shall outline what is needed to make high school mathematics functional, the core of which is the teaching of modelling; modelling also reinforces the learning of mathematical concepts and skills (see Burkhardt and Muller 2006) This is not a zero-sum game.

**Skill in the modelling process** is a key component in 'doing mathematics'. The diagram (see, e.g., Burkhardt 1981) shows a standard outline of its key phases [7]. In current mathematics assessment and teaching, only the SOLVE phase gets much attention. (The situation is sometimes better in Statistics curricula)

Mathematical modelling is not an everyday term in school mathematics; indeed, it is often thought of as a rather advanced and sophisticated process, used only by professionals. That is far from the truth; we do it whenever we *mathematize* a problem. The following tasks illustrate this:

- Joe buys a six-pack of coke for £3 to share among his friends. How much should he charge for each bottle?
- If it takes 40 minutes to bake 5 potatoes in the oven, how long will it take to bake one potato
- If King Henry 8th had 6 wives, how many wives had King Henry 4th?

The difference between these tasks is in the appropriate choice of mathematical model. The first is a standard proportion task; however, *all the tasks in most units on proportion need proportional models*, *so skill in choosing an appropriate model is not developed*. In the second case, it depends on the type of oven (constant = 40 mins for traditional *constant temperature* ovens; roughly proportional = 40/5 = 8 mins for *constant power* microwave ovens) For each problem, as usual, more refined models could also be discussed – the <improve> choice in the EVALUATE phase in the diagram. On the third task, if they laugh they pass.




Mathematics teachers sometimes argue that choosing the model is "not mathematics", but it is essential for mathematics to be functional. Of course, the situations to be modelled in

[7] The phases of pure mathematical problem solving are similar.

mathematics classrooms should not involve specialist knowledge of another school subject but, as with the above examples, situations that children meet or know about from everyday life. English teachers reap great benefits from bringing students' lives into their teaching; where mathematics teachers have done the same (see, e.g., Shell Centre 1987-89), motivation is improved, particularly but not only with weaker students, and relationships in their classrooms are transformed[8]. Mathematics acquires human interest. Curriculum design is not a zero-sum game; the use of 'math time' in this way enhances students' learning of mathematics itself (Burkhardt and Muller 2006).

**What mathematics content should we include?** There will always be diverse views on this. This is not the place to enter into a detailed discussion of what mathematical topics should have what priority. (For this see, e.g., National Research Council 2001.) Here I shall only raise a few aspects of US curricula that, from an international perspective, seem questionable. Is a year of Euclidean Geometry a reasonable, cost-effective use of every high school graduate's limited math time, or should it be specialist mathematics – an extra option for enthusiasts? Should not the algorithmic/functional aspect of algebra, including its computer implementation in spreadsheets and programming, now play a more central role in high school algebra? (Mathematics everywhere is now done with computer technology – except in the school classroom.) Should calculus be a mainstream college course, to the exclusion of discrete algorithmic mathematics and its many applications, or one for future specialists in the physical sciences and traditional engineering?

In the UK, there is high-level move (Smith Report 2004, Tomlinson Report 2004) to address such issues by introducing "double mathematics" from age 14, with a challenging functional 'mathematics for life' course for all and additional specialist courses with a science and engineering, or business and IT focus. It will be interesting to see how this develops. (We already have English Language and English Literature. All students take the first; about half take both)

## Dimensions of performance in Mathematics

Whenever curriculum and assessment choices are to be made, discussion should focus on performance as a whole, not just the range of mathematical topics to be included. To support such an analysis, MARS has developed a *Framework for Balance*, summarised in the table below (Balanced Assessment 1997–99 contains an outline). You will note, as well as the familiar *content* dimension, the *phases of problem solving* from the previous diagram, and various others including one holistic dimension – *Task Type*. This multi-dimensional analytic scheme (it is dense, and takes time to absorb) allows one to check balance in all the major dimensions of performance. In most current tests, balance is sought only across the content dimension, and the only task type is short *exercises* that require only the SOLVE ~ transformation and manipulation phases[9]. Formulation is trivialised, while interpretation, critical evaluation and communication of results and reasoning are rarely assessed in mathematics tests.

[8] "The Three R's for education in the 21st century are Rigour, Relevance and Relationships", Bill Gates, US National Governor's Conference 2005. Functional mathematics develops them all.

[9] The common argument that "You need a solid basis of mathematics before you can do these things" is simply untrue. However small or large your base of concepts and skills, you can deploy it in solving worthwhile problems - as young children regularly show, using counting. Deferring these practices to graduate school excludes most

**Task types** I have space here to illustrate only this holistic dimension of the otherwise-analytic *Framework for Balance*, and that with only one task of each type. I choose it because it brings out something of the variety of challenges that mathematics education and assessment should aim to sample (cf literature, science, social studies, music ....) Tasks are mostly given here in their core form; for any specific grade, they need to be appropriately engineered. The design goal is to enable *all* students who have worked hard in a good program to make significant progress, while offering challenges to the most able. This can be achieved in various ways including 'open tasks' or 'exponential ramps' to greater generality, complexity and/or abstraction.

---

people, and stultifies everyone's natural abilities in real problem solving . It is also an equity issue – such deferred gratification increases the achievement gap, probably because middle class homes have time and resources to encourage their children to persist in school activities that lack any obvious relevance to their current lives.

# Framework for Balance

*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.

  **Other Mathematics** including: discrete mathematics, including combinatorics; underpinnings of calculus; mathematical structures.

*Mathematical Process Dimension*

- **Phases** of problem solving, reasoning and communication will include, as broad categories, some or all of:

  Modeling and Formulating;
  Transforming and Manipulating;
  Inferring and Drawing Conclusions;
  Checking and Evaluating;
  Reporting.

*Task Type Dimensions*

- **Task Type** will be one of: open investigation; non-routine problem; design; plan; evaluation and recommendation; review and critique; re-presentation of information; technical exercise; definition of concepts.
- **Non-routineness** in: context; mathematical aspects or results; mathematical connections.
- **Openness**: it may have an open end with open questions; open middle.
- **Type of Goal** is one of: pure mathematics; illustrative application of the mathematics; applied power over the practical situation.
- ***Reasoning Length*** *is the expected time for the longest section of the task (It is an indication of the amount of 'scaffolding' – the detailed step-by-step guidance that the prompt may provide)*

*Circumstances of Performance Dimensions*

- **Task Length**: ranging from short tasks (5-15 minutes), through long tasks (15-60 minutes), to extended tasks (several days to several weeks)
- **Modes of Presentation**: written; oral; video; computer.
- **Modes of Working** on the task: individual; group; mixed.
- **Modes of Response** by the student: written; built; spoken; programmed; performed.

Of the following two ***planning tasks***, note that the second is more open, giving less specific guidance.

### *ICE CREAM VAN*

*You are considering driving an ice cream van during the Summer break. Your friend, who "knows everything", says that "it's easy money". You make a few enquiries and find that the van costs £60 per week to hire. Typical selling data is that one can sell an average of 30 ice creams per hour, each costing 50c to make and each selling for $1.50.*

*How hard will you have to work in order to make this "easy money"?*

### *TIMING TRAFFIC LIGHTS*

*A new set of traffic lights has been installed at an intersection formed by the crossing of two roads. Right turns are NOT permitted at this intersection.*

*For how long should each road be shown the green light?*

A study (Treilibs et al 1980) of the responses to these tasks of 120 very high-achieving Grade 11 mathematics students found that *none* used algebra for the modelling involved. (They used numbers and graphs, more or less successfully). Yet they all had five years of successful experience with algebra but, with no education in real problem solving, their algebra was non-functional. Modelling skill is important and, as many studies (e.g. Shell Centre 1987-89) have shown, teachable.

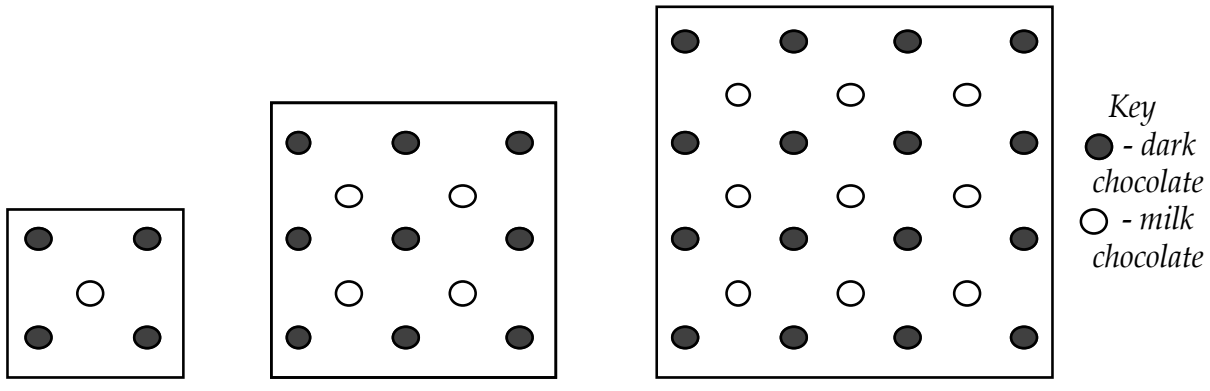
The next task is typical of a genre of ***non-routine problems*** in pure mathematics, often based on pattern generalisation, in which students develop more powerful solutions as they mature.

### *SQUARE CHOCOLATE BOXES*

*Chris designs chocolate boxes.*

*The boxes are in different sizes.*

*The chocolates are always arranged in the same kind of **square** pattern.*
*The shaded circles are dark chocolates and the white circles are milk chocolates.*




*Chris makes a table to show how many chocolates are in each size of box.*

| *Box number* | *1* | *2* | *3* | *4* | *5* |
|---|---|---|---|---|---|
| *number of dark chocolates* | *4* | *9* | | | |
| *number of milk chocolates* | *1* | *4* | | | |
| *total number of chocolates* | *5* | *1* | | | |

*1. Fill in the missing numbers in Chris's table.*
*2. How many chocolates are there in Box 9? Show how you figured it out.*
*3. Write a rule or formula for finding the total number of chocolates in Box* n*.*
*Explain how you got your rule.*
*4. The total number of chocolates in a box is 265. What is the box number?*
*Show your calculations.*

The scaffolding shown for this task (MARS 2000– ) fits the current range of performance in good middle grade classrooms. One would hope that, as problem solving strategies and tactics become more central to the curriculum, #3 alone would be a sufficient prompt. The following is an ***evaluate and recommend*** task – an important type in life decisions, where mathematics can play a major role.

### *WHO'S FOR THE LONG JUMP?*

*Our school has to select a girl for the long jump at the regional championship*
*Three girls are in contention.*
*We have a school jump-off. These are their results, in meters*

| *Elsa* | *Ilse* | *Olga* |
|---|---|---|
| *3.25* | *3.55* | *3.67* |
| *3.95* | *3.88* | *3.78* |
| *4.28* | *3.61* | *3.92* |
| *2.95* | *3.97* | *3.62* |
| *3.66* | *3.75* | *3.85* |
| *3.81* | *3.59* | *3.73* |

*Hans says "Olga has the longest average.*
*She should go to the championship"*
*Do you think Hans is right? Explain your reasoning.*

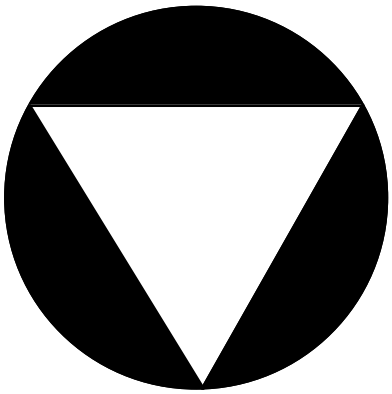
This task provides great opportunities for discussing the merits and weaknesses of alternative measures, here in the context of track and field long jump. However, in the TIMSS video lesson on which this task is based (from Germany, but it could be in the US), the students calculate the mean length of jump for each girl and use that for selection. (Olga wins, despite having shorter longest jumps than either of the others) The teacher moves on without comment! A splendid opportunity is missed – to discuss other measures, their strengths and weaknesses, the effect of a "no jump", or any other situational factors. (Poor Bob Beamon – 3 no jumps, then a very long-

lasting world record. He's out.) Is this good mathematics? I have found research mathematicians who defend it as "not wrong". What does this divorce from reality do for the image of mathematics?

*Magazine Cover* is a ***re-presentation of information*** task (for Grade 3, but adults find it non-trivial) It assesses geometry and mathematical communication.

*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.*

The rubric (below, from MARS 2000– ) illustrates how complex tasks can, with some scorer training, be *reliably* assessed – as is the practice in most countries and, locally, in AP exams.

| | **Magazine Cover: Grade 3** | Points | Section points |
|---|---|---|---|
| | *Core elements of performance:* • *describe a geometric pattern*<br>*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 corne**r. *accept*: Upside/going down. | 1 | |
| | Describes measurements of circle/triangle. | 1 | |
| | Describes color: **black/white.** | 1 | |
| | **Allow 1 point for each feature up to a maximum of 6 points.** | | |
| | **Total Points** | | 6 |

For our last example, we return to a type that, perhaps, best represents 'doing' both mathematics and science – ***investigation***. *Consecutive Addends* (above) is an open investigation in pure mathematics. Equally, there are many important situations in everyday life that merit such investigation. One important area, where many children's quality of life is being curtailed by their parents' (and society's) innumeracy, is tackled in:

### *BEING REALISTIC ABOUT RISK*

*Use the web to find the chance of death each year for an average person of the same age and gender as*

- *you*
- *your parents*
- *your grandparents*

*List the things that people fear – being:*

- *struck by lightning*
- *murdered*
- *abducted by a stranger*
- *killed in a road accident*
- *winner of the lottery*
- *……..*

*For each, find out the proportion of people it happens to each year.*

*Compare real and perceived risks and, using this information, write advice to parents on taking appropriate care of children.*

There will need to be more emphasis on ***open investigations***, pure and real-world, if the quality of mathematics education, and students' independent reasoning, is to improve.

The above tasks, and the Framework for Balance, provide the basis for a response to our question *What mathematics values should assessment reflect?* Taken together, they give a glimpse of the diversity of assessment tasks that enable students to show how well they can do mathematics – "making music" not just "practising scales". There is a place in the *Framework for Balance* for ***technical exercises*** too – but even these don't have to be boring:

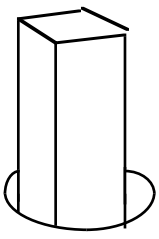
### *SQUARE PEG*

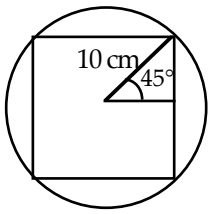
*Lee has heard of an old English proverb used when someone is doing a job that they are not suited to.*

*The proverb describes the person as "fitting like a square peg in a round hole".*

*Lee wondered how much space was left if a square peg was fitted into a round hole.*

*Lee constructed a square that just fitted inside a circle of radius 10 cm*






*1. What percentage of the area of the circle is filled by the area of the square?*

*Explain your work and show all your calculations.*

Part 2 of this task asks for the same calculation for a circle inside a 10cm square hole.

Published examples of such tasks include: (MARS 2000–), a set of annual tests for Grades 3 through 10, New Standards (1998) released tasks, and Balanced Assessment (1997-99), classroom materials for assessment and teaching. *World Class Tests* (MARS 2002-04) provides a more challenging range of tasks, aimed at high-achieving students,

## Improving quality in assessment design

Designing and developing good assessment tasks, which have meaning to students and demand mathematics that is important for them, is among the most difficult educational design challenges. The tasks must enable students *to show what they know, understand and can do* without the help from teachers that classroom activities provide. Task design is usually subject to too-tight constraints of time and form. Starting with a good math problem is necessary, but far from sufficient. As in all design:

**Good design principles are not enough; the details matter**[10]

Thus it is important to recognize high-quality in assessment tasks, and to identify and encourage the designers who regularly produce outstanding work. The latter are few and hard to find. (Swan et al 1985) contains some well-known benchmark examples.

The emphasis in this paper on the task exemplars is no accident, but it is unconventional; without them, the analytic discussion lacks meaning. In a misguided attempt to present assessment as more 'scientific' and accurate than it is, most tests are designed to assess elements in a model of the domain, which is often just a list of topics. All models of performance in mathematics are weak, usually taking no account of how the different elements interact.

*Our experience with assessment design shows that it is much better to start with the tasks.* Get excellent task designers to design and develop a wide range of good mathematics tasks, classify them with a domain model, then fill any major gaps needed to balance each test.

Interestingly and usefully, when people look at specific tasks, sharply differing views about mathematics education tend to soften into broad agreement as to whether a task is worthwhile.

*"Yeah, our kids should be able to do that"*

Having looked in some depth at the values that underlie performance, we now have the basis for answering the other questions with which I began. I shall be brief and simplistic.

*Who is assessment for? What is it for?* Governments, and some parents, want it for accountability. Universities and employers for selection. They all want just one reliable number. Teachers and students, on the other hand, can use a lot of rich and detailed feedback to help diagnose strengths and weaknesses, and to guide further instruction. Some parents are interested in that too.

*When should it happen to achieve these goals?* For teachers and students in the classroom, day-by-day – but, to do this well[11], they need much better tools. For accountability, tests

[10] The difference between Mozart, Salieri, and the many other composers of that time we have never heard of was not in the principles – the rules of melody, harmony, counterpoint, and musical form. Students deserve tasks with some imaginative flair, in mathematics as well as in music and literature.

[11] The classroom *assessment for learning* movement is relatively new. There is much to do.

should be as rare as society will tolerate; the idea that frequent testing will drive more improvement is flawed. Good tests, that *will* drive improvements in curriculum, need only happen every few years.

*What will the consequences be?* Because effective *support* for better teaching is complex and costs money while *pressure* through test scores is simple and cheap, test-score-based sanctions seem destined to get more frequent and more severe. The consequences for mathematics education depend on the quality of the tests. Traditional tests will continue to narrow the focus of teaching, so learning, which relies on building rich connections for each new element, will suffer. Balanced assessment will, with some support for teachers, drive continuing improvement. Currently, with the air full of unfunded mandates, the chances of improved large-scale assessment do not look good.

## Cost and cost-effectiveness

Finally, *What will it cost, and is this an appropriate use of resources?*

*Feedback* is crucial for any complex interactive system. Systems that work well typically spend ~10% turnover on its 'instrumentation'. In education total expenditure is ~ $10,000 per student-year which suggests expenditure of **~ $1,000 per student-year on assessment** across all subjects. Most of it should be *assessment for learning* in the classroom[12], with about 10%, **~$100 a year on summative assessment** linked to outside standards. This is an order of magnitude more than at present but still only 1% of expenditure. Increases will be opposed on all sides for different reasons, particularly budget shortage by administrators and dislike of assessment by teachers. Yet while 'a dollar a student' remains the norm for math assessment, students' education will be blighted by the influence of narrow tests[13].


## *Acknowledgements*

Malcolm Swan and Rita Crust led the design of many of these tasks, which were developed and refined in classrooms in the UK and the US by the MARS team. I have been fortunate to work with them all.


## *References*


Balanced Assessment (1997-99) *Balanced Assessment for the Mathematics Curriculum*, eight volumes of classroom assessment, Parsippany, NJ: Dale Seymour Publications, Pearson Learning.

Burkhardt, H. (1981) *The Real World and Mathematics,* Glasgow, UK: Blackie-Birkhauser; reprinted 2000, Nottingham, U.K.: Shell Centre Publications, http://www.mathshell.com/scp/index.htm

Burkhardt, H. and Muller, E. (2006) *Applications and Modelling for Mathematics* in Blum, W. and Henn, H-W. (Eds) *Applicaions and Modelling in Mathematics Education,* ICMI Study 14. Amsterdam: Kluwer.

Crowther Report (1959) 15-18: *A Report of the Central Advisory Council for Education*, London: HMSO.

MARS: Crust, R., Burkhardt H. and the MARS team (2000– ) *Balanced Assessment in Mathematics*, annual tests for Grades 3 through 10, 2001-2004 Monterey CA: CTB/McGraw-Hill, East Lansing, 2005– MI: MARS, http://www.math-shell.com/scp/index.htm

MARS: Swan, M., Crust, R., and Pead, D., with the Shell Centre team for the UK Qualifications and Curriculum Authority (2002-4) *World Class Tests* of *Problem Solving in Mathematics Science and Technology,* London: nferNelson.

New Standards (1998) *New Standards Mathematics Reference Examination*, San Antonio, TX: Harcourt Assessment


[12] Professor, on seeing abysmal student scores a third of the way through his analysis course: "We've gone so far in the semester, I don't know what to do except to go on – even though it's hopeless."

[13] If, for reasons of economy and simplicity, you "judge the decathlon by running only the 100 meters", you may expect a distortion of the training program.

National Research Council (2001) *Adding it Up: Helping Children Learn Mathematics*. Washington, DC: National Academy Press.

PISA (2003) *The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills*, Paris: OECD, https://www.pisa.oecd.org/dataoecd/38/51/33707192.pdf

Shell Centre: Swan, M., Binns, B., & Gillespie, J., and Burkhardt, H, with the Shell Centre team (1987-89) *Numeracy Through Problem Solving,* Longman, Harlow; reprinted 2000, Shell Centre Publications, Nottingham, U.K., URL: http://www.mathshell.com/scp/index.htm

Smith Report (2004) *Making Mathematics Count*, UK Department For Education and Skills, London: HMSO, http://www.mathsinquiry.org.uk/report/index.html

Steen, L. A.(ed): 2002,*Mathematics and Democracy: the case for quantitative literacy,* Washington, DC: National Council on Education and the Disciplines, http://www.maa.org/ql/mathanddemocracy.html

Swan, M., Pitts, J., Fraser, R., and Burkhardt, H, with the Shell Centre team: 1985, *The Language of Functions and Graphs,* Manchester, U.K.: Joint Matriculation Board, reprinted 2000, Nottingham, U.K.: Shell Centre Publications, http://www.mathshell.com/scp/index.htm

Tomlinson Report: 2004, *14-19 Curriculum and Qualifications Reform,* Department For Education and Skills, London: HMSO, http://www.14-19reform.gov.uk/

Treilibs, V., Burkhardt, H. & Low, B.: 1980, *Formulation processes in mathematical modelling*, Nottingham: Shell Centre Publications, http://www.mathshell.com/scp/index.htm