ASSESSING GIFTED AND TALENTED CHILDREN
World Class Tests International Seminar – February 2001
Chapter for the QCA Book

World Class Assessment: Principles, Practice and Problem Solving

Hugh Burkhardt
MARS = Mathematics Assessment Resource Service
Shell Centre, School of Education
University of Nottingham and Michigan State University

The assessment of problem solving epitomises all the problems of designing and developing high quality assessment. Indeed, any test that goes beyond the routine, covering more than learned facts and procedures in familiar contexts, assesses problem solving – in some sense of that rather-too-widely used phrase. Most assessment makes that claim, though often the tasks the students are asked to do and the aspects of performance rewarded in the scoring schemes do not match those claims. Similarly, the assessment of gifted and talented children simply highlights more general problems. Clearly mundane and narrow assessment tasks are not good enough for them; neither should they be for any child.

To summarise the essentials of problem solving:

- a 'problem' is non-routine, in some sense an unfamiliar task;
- the solver must find, not just remember, a solution path.

So beyond 'the basics', everything is problem solving, though the degree of unfamiliarity (the*transfer distance*) will vary. The theme of this chapter will be the central importance of:

- the richness of the tasks on which students are assessed;
- the range, variety and balance of the task set, sampled in each test.

With this as a theme, examples are essential. Those presented here are mostly from MARS work on developing tasks, for World Class Tests and for other projects. Here they are simplified to save space, sometimes by removing some of the 'scaffolding' which helps give students access to the problem. For these tasks and many other things I am indebted to the MARS team, here particularly Malcolm Swan, Jim Ridgway, Rita Crust and Alan Bell. (More about MARS' work, including references, can be found at www.educ.msu.edu/MARS/)

The examples will run parallel to the text. They are mainly from the area where we have most experience – mathematics in the broad sense, with an emphasis on solving substantial problems, many involving the application of mathematics in practical situations. This is the now-standard view of mathematics in the international community and in the UK National Curriculum, at least in principle – in practice, various assessment and curriculum pressures have produced sharp narrowing. The examples here are all of paper-based tasks; I am sorry that I can't show some of the computer-based tasks, which are one of the exciting features of the World Class Tests project. (You can find some examples on the QCA website www.qca.org.uk/fsmu/).

Design a Tent is a good example to start with. It is a mathematics task, and a problem solving task, with a strong flavour of design technology. It is a problem of obvious practical relevance. Even if few students will become tent designers, this is the kind of design task that we all need to be able to think through – if only to take a critical look at designs with which we are presented. For the moment, please just work out how you would do it; we shall comment further later.

[Insert Design a Tent near here]

Assessment Tasks

The design and development of good assessment tasks is among the most challenging aspects of educational design. The teacher observation and targeted guidance that characterises good lessons cannot be used. The tasks have to "work on their own", allowing students *"to show what they know, understand and can do*" across the domain..

I will outline some principles for designing and developing high-quality assessment, some of the difficulties that arise in practice, and how the degradation they produce may be minimised. This may suggest familiar ground, and I hope that much of what I say will indeed seem obvious. However, I choose it because the quality of so much of the assessment that children face around the world is undermined by neglecting the "obvious". There are often "sound, practical reasons" adduced for this: "We don't know how to do that", "It won't be reliable enough", "It's all too complicated", "It will cost too much", and so on.

These are real challenges but assessment design, like good design in any field, is about finding attractive ways around such obstructions within the constraints that every design problem involves. Modern airplanes, for example, are significantly more comfortable, faster and safer, than the Wright brothers' prototype though all the objections just listed applied there. As in every field, we can do better.

In MARS' work for all its client systems, we are seeking high quality in this sense. In developing World Class Tests, there are some interesting new challenges. MARS experience is largely in Mathematics; the assessment of "Problem Solving in Mathematics, Science and Technology" presents even broader challenges, of task design and of domain definition. We are making progress with these but there is much still to learn. The same can be said about the design and marking of computer-based tasks though, because of the usual complexities of software development, we are not yet as far forward there. I mention these two exciting aspects of World Class Tests to make a more general point – that the design of assessment for problem solving is not a routine procedure. The design of assessment of skills beyond the routine cannot itself be routine.

Now to the principles.

The Validity Principle – assess what you actually want 'them' to be able to do.
What could be simpler or more obvious? The fundamental criterion for a good task is that a well-educated person should say:

"Yes, that is the kind of thing our kids should be able to do"

Pollen is such a task, involving data analysis and scientific inference. Such data-based tasks often also involve *evaluation and recommendation*, with the student in a *consultant role*.

[Insert Pollen near here]

Yet, in most assessment, such face validity is hard to find. Instead, you find tasks that imitate the exercises that students are given in the curriculum – a very different thing. Why? Neglect of the next principle is a major reason.

The Holistic Principle – assess samples, not ingredients, of performance
This links closely to the Validity Principle. It is widely ignored for the following attractive and plausible reason. "Performance is made up of a limited set of separate 'ingredients'; if we assess each of these, we have assessed performance." These ingredients are sometimes called detailed behavioural objectives. In the original National Curriculum in Mathematics they were called Statements of Attainment. The idea is to assess whether each student can, or can not, do each of them. Simple and appealing – and not obviously stupid.

Unfortunately, it doesn't work for two major reasons. If the ingredients are to have a specific difficulty level, it turns out you need to specify them in great detail – not "can multiply whole numbers" but "can multiply a 3-digit number by a 2 digit number, set out in standard form, with carrying involved". So you need thousands of separate "assessment criteria", which becomes impracticable. However this is a relatively minor problem, compared to the "composite task" effect.

For a substantial task that involves a number of such elements of performance, the whole is much more than the sum of the parts. Assessing a student on each bit separately tells you little about how they would do on the task as a whole. If you want to see how good someone is at driving a car, would you just test them separately on: starting the engine; turning the steering wheel; changing gear; letting in the clutch; signalling, ... – or would you go with them for a drive. Not only does performance on each ingredient *not* tell you how well they can drive, but assessing their driving as a whole *does* show whether they can do all the essential elements.

Design a Tent would be far less valuable or plausible task if it was broken up into a sequence of steps such as: Estimate how tall a person might be. How much space should you leave for their baggage? How wide a sleeping space should you leave for each person? How tall should the tent be so they can move around kneeling? Work out are the dimensions of the base rectangle? Use the Pythagorean Theorem to calculate the height of the sloping sides, and the dimensions of the triangular ends. Sketch the shape of the canvas 'top' of a tent, showing these dimensions.....

The essential point is to assess whether the student can handle the strategy as well as the techniques needed to solve the problem – can choose which tools to use, as well as using them reliably (like climbing a mountain without, or with, an expert guide).

The problems we face in life and work are often substantial and come without a guide. So without experience of solving unfamiliar substantial problems, our skills are of limited use in

practice. (When did you last use the Pythagorean Theorem, or find the roots of a quadratic? Yet they can be useful.) The National Curriculum in Mathematics recognises this fully in its principles but, though Statements of Attainment are long gone, the fragmentation they prescribed still dominates. To check on this, just look at the length of most of the assessment tasks – or rather, because they are often broken into many small parts, the length of the *longest part* of each task (which we call the *reasoning length*). For most mathematics assessment tasks, this is only a minute or two; how many significant problems we meet in everyday life are so short?

Encouraging curriculum balance – the WYTIWYG Principle

With high-stakes assessment, where the results have important consequences for those involved, *What You Test Is What You Get* in the classroom. So balanced assessment is a key principle for MARS (indeed it is the title of our first international project). It defined it as the assessment designers' responsibility to ensure that teachers who "teach to the test" are led to provide a rich and balanced curriculum.

Teachers naturally emphasize those aspects of performance that are "on the test". They would be very brave (and perhaps irresponsible) not to, since society decrees that their students and their performance be judged by the results. So, in the language of TIMSS, if the test does not cover the performance goals of the *intended curriculum*, the actual *implemented curriculum* will narrow – essentially, to match the *tested curriculum*. In this way unbalanced high-stakes assessment always distorts the curriculum; all those who perpetrate it carry that responsibility.

In practice, it often happens 'by accident', through lack of awareness. Thus when the Dearing Report, in response to pressure, recommended the simplification of the Key Stage tests, they also said that the missing areas should be covered by teacher assessment in the classroom, which should "carry equal weight". This second aspect was not implemented.

Problem solving, in any subject, is a frequent casualty of narrowing. Some try to justify it: "You can't (or "It's not fair to) assess problem solving in a timed exam." But comparing students' performances on exam tasks, taking between 5 and 20 minutes, with other assessments shows that it is possible to do it fairly. *Snakes and Ladders*, the next example, is a simple mathematical task that assesses important aspects of the design process, particulrly critique and improvement.

[Insert Snakes and Ladders near here]

The importance of balance is such that I will reinforce the argument with an example from another field of assessment. Suppose your goal is to develop athletes for the decathlon, with the broad and balanced skills that this event requires. They train on ten events and, of course, you want to assess them on all these. But this takes two days, a big stadium (for running the 1500 metres, throwing the discuss etc) with lots of equipment and officials. People may say "It's too complicated" "We don't want to waste time on assessment; it's the training that matters." "Let's keep it simple." "How can we do this? Well, performance on the 100 metres is simple to measure. It's well correlated with general athletic ability." (Good all-round athletes are usually pretty good at it, much better than the unathletic) "Great idea – we'll decide the winner of the decathlon with just a 100 metres race." Do you think this might have some effect on the balance

of training? This kind of absurdity happens unnoticed in many kinds of assessment – the tests students actually take cover only a small part of the range of things they are supposed to be learning. Soon the implemented curriculum shrinks to match the test.

However, the balance principle is not just a challenge but also an opportunity. For various reasons, not simply because of narrow assessment, the curriculum in many classrooms is far narrower than the intended curriculum. When new types of task are introduced into high-stakes assessment, such as GCSE, A-level or the National Tests, teachers will try to bring those new aspects of performance in the subject into their teaching. If they are to do it well, many will need help in the form of well-aligned teaching materials and professional development support. We at the Shell Centre in Nottingham developed this model in the 1980s with the JMB for its O-level, introducing one new task type each year. (*Problems with Patterns and Numbers*, *The Language of Functions and Graphs*, Shell Centre, 1984, 1986) This model worked well, and the materials still sell, but it was swept away, with the JMB, by the introduction of the GCSE. Others including the National Numeracy Strategy have used a similar approach.

What is balance?

This is a huge topic, raising all kinds of issues, which I do not have space to go into here. Ultimately balance is a judgment for the professional peer group in the subject, guided by society at large. However, it is worth outlining the kind of analytic framework within which balance should be explored, for it is here that an assessment domain is defined.

I have emphasised the importance in designing tasks of a holistic approach, focussed on worthwhile tasks rather than the separate elements of performance. But these elements are part of an analytic framework for defining the domain of assessment. It is in looking at balance that the analytic and holistic best come together. "OK, so this is an attractive collection of tasks. Can you describe what they assess?"

The essential feature of a framework is that it should address *all* the aspects of performance that may be regarded as important. For this it must be multidimensional. A driving test must address the *technical* things we have mentioned (steering etc). It must also be concerned with other *strategic* dimensions of performance (awareness of other road users, choosing appropriate actions, even navigation) and how they all interact. The *Dimensions of Balance* in Table 1 summarise the framework MARS has developed for Mathematics over the last decade. (A parallel framework for "Problem Solving in Mathematics, Science and Technology" is now being developed by MARS, with QCA and international experts, as part of the World Class Tests) The first point of note is that the "Content" of concepts and skills is only one important dimension; to be balanced across content, which most tests are, is not enough to claim that one assesses performance in mathematics. What of the other dimensions of performance? Brief comments must suffice here. (A detailed handbook, illustrated with many tasks, is available for those who want to know more, *Bell, Burkhardt et al, 1996, revised 2000*)

[insert Table 1 near here]]

Task length and *Reasoning length* have already been mentioned. Balance here is clearly important; if you want people to be able to sustain chains of reasoning longer than a minute or two, you had better include an appropriate proportion of tasks that require it.

Process dimensions address the strategic and tactical aspects of performance – the exploring, planning, carrying through, reflecting on, and communicating the solution to the task. These major *Phases* are all essential elements in performance and should be substantially represented in any balanced test. Most mathematics assessment is unbalanced in this regard, with an overwhelming proportion of the *transforming and manipulating* aspect. (Better balance requires longer tasks) One can analyse process aspects in much more detail; we do so for research purposes but it does not seem essential for balancing tests, where checking on all the dimensions is more important than going into great detail on any one.

The other dimensions also need attention in balancing. *Goal type* and *Context* dimensions, for example, are there to identify the "using and applying of mathematics".

[Insert Rope and Hexcube near here]

Finally,*Task type* is a useful dimension of a slightly different kind, summarising combinations of the other aspects within a single task in a way that is easy to understand. The types listed represent important ways of doing mathematics. We have already seen examples of Design and Plan (*Design a Tent* and *Snakes and Ladders*) and Re-present Information (*Pollen*), which includes inferring the meaning in the problem context. *Rope* is another of this important type, while *Hexcube* is a non-routine problem in geometry, with some possible design implications. Below we shall show an Open Investigation, *Consecutive Sums*, and an Evaluate and Recommend task, *Crown.* Range and variety of task type is central to balancing any test.

Our draft framework for Problem Solving has all these dimensions, but with definitions broadened to cover those aspects of doing Science and Technology that are not much found in Mathematics. I hope this brief outline brings out the constructive dialogue between the holistic and analytic viewpoints – the task set and the domain framework for balance. Both are essential but I believe that, as in most fields, the holistic should take priority. (You can infer the rules of melody, harmony and counterpoint from The Marriage of Figaro; you cannot deduce that marvellous creation from the rules)

Finally, let me briefly mention four more principles that should not be ignored.

The Learning Principle – assessment time should be learning time, too

The traditional "measurement" view of assessment is that it should assess what the student knows, understands and can do with as little cost as possible in time and money. It has no other function; above all, it has nothing to do with learning. As well as being false, this view also represents a tremendous missed opportunity. In a world where the classroom activities are not always ideal stimulants to learning, good high-stakes assessment can be a powerful *engine for improvement*. It is (unfortunately) possible that in many classrooms, some good assessment

tasks will be among the most interesting and stimulating learning activities the student meets. It is certainly a worthy design ambition. Further, because they command the attention of students and teachers, they will be taken seriously.

[insert Consecutive Sums near here]

Investigative microworlds provide one excellent genre. (The computer offers a fine environment for presenting these, both in mathematics and science) *Consecutive Sums* is one such – a simple but rich mathematical microworld with a host of interesting results to discover, and to explain, in open investigation. For example, you will find by exploring examples that the powers if 2 (2, 4, 8, ...) can not be expressed as sums of consecutive whole numbers. Why? This is a challenging question, but there are other interesting easier ones too.

The Literary Value Principle

In the assessment of language, it is taken for granted that the texts that students face should be better-than-mundane. They should have value as pieces of writing worth reading. There may be argument over which pieces are appropriate but the principle of literary value is accepted. The same cannot be said of most assessment tasks in mathematics or science. Do they embody something memorable in the subject? Do they spark any interest in the students? Rarely – they are usually profoundly mundane. Could we do better?

Perhaps *Crown* is a rather obvious example but it and some of the other tasks here suggest something of what can be done.

[insert Crown near here]

The Reliability Principle – know it, but don't make it the priority

The results of assessment must have meaning, so reliability is important. I mention it to show it is not forgotten; it is simply not the main theme here. I would say, however, that it is dangerous to make accuracy of measurement a priority. It so often leads to violating the other principles above – to measuring those things that are easy to assess, justifying it by correlation arguments. Height is well-correlated with mathematical performance over the age range 5 to 18. Why not just measure that?

The Design Principle – simplicity in use may mean sophistication in design

I hope this needs no elaboration.

Automata or Thinkers?

Coming full circle, I want to return to focus on what our students need from the curriculum and its assessment. Why is problem solving so important? At one time, a person with a limited set of handwriting or calculating skills could rely on them to make a fair living; now there are very inexpensive *automata* (calculators, word processors, ...) that perform all these functions reliably. They cost anything from a few pounds for a basic calculator to less than £100 for something that will do all the technical operations of school mathematics (graphing, statistics, algebra,...). The

mathematical education of a student from age 5 to 18 costs roughly £3,000! What is society, and the student, getting for the money?

In life and work people are now asked to tackle non-routine tasks which they have not practised at school. If they are to be employable, they need a much wider range of higher level strategic skills – ie problem solving, including skills, reasoning, organisation and communication. The need is for flexibility, adaptability, and self-propelled learning.

People who are trained simply to be efficient *automata* are losing their jobs all over the world to real robots. Our economic prosperity depends on being a nation of *thinkers*. These are the attributes in which humans can excel; they also bring personal satisfaction. More flexible, adaptable thinkers are what a country needs. This applies particularly to those countries with a high standard of living like ours. As Marc Tucker of New Standards put it:

> *The choice for us is between a high-skill high-wage or a low-skill low-wage economy; we can't for much longer enjoy a low-skill high-wage economy*

We must ask, subject by subject, how the different elements of curriculum and of assessment are contributing to preparing us for this change. It seems clear that gifted and talented children, who cope easily with the current curriculum, should be educated to this broader range of challenges. We hope that World Class Tests will provide both stimulus and, through the surround materials, some support for this.

There are wider implications for all children – one must surely also ask whether the ability to solve varied problems, an essential complement to skill acquisition at any level, needs more attention in both curriculum and assessment.