Common failures in strategic design

These two examples, one from assessment design and one from the design of curriculum documents, are from my forthcoming paper for "On Strategic Design". Comments welcome. Hugh Burkhardt

Assessment – the "only measurement" fallacy

Policy makers in the anglophone countries and some others are wedded to using tests of various kinds as prime instruments of system control Tests are seen as reliable measures[1] of student, teacher and school performance, forming the basis of each school's "accountability" to the society that funds it. Targets are often set in terms of test scores so these have serious consequences for those concerned. Students access to higher education depends on their test scores. In England, schools are ranked on test scores into "league tables" to guide "parental choice[2]". Schools that under-perform may be "taken into "special measures" or closed.

Given the importance of tests, it seems obvious that their design should be a focus of attention. They should embody the full set of performance goals in a balanced way[3]. Yet this central responsibility of test providers and those that commission test design is widely ignored, and sometimes denied. The focus is on the statistical properties of the test and the "fairness" of the procedures, with little attention to what aspects of performance are assessed[4]. Policy makers talk and behave as though tests are just "measurement"[5]. They choose simple tests because they are cheap and, if pressed, argue that they are adequate are believed to correlate with more valid and elaborate assessments.

However, this approach ignores two of the three roles that high-stakes assessment *inevitably* plays – it:

A. *measures levels of student performance,* but only across the range of task-types used;
B. *exemplifies performance objectives* – the types of task in high-stakes tests show, in a clear form that teachers and students readily understand, what kinds of performance will be recognized and rewarded; as a result, this set of task types
C. *dominates classroom activities* – there is plenty of evidence (some of it reported below) that the task types in high-stakes tests largely determine the pattern of teaching and learning activities in most classrooms.

Thus assessment design is the unnoticed "elephant in the room". There is plenty of evidence that "what you test is what you get" *(WYTIWYG)* is a fact of life – in systems with high-stakes assessment, the tests *are* the *de facto* standards. The UK national inspectors of schools (Ofsted 2006, 2008) remark on the dominance

---

[1] "Goodhart's Law" states that "When a measure becomes a target, it ceases to be a good measure" – essentially because targets promotes gaming and other distortions described here. Dylan Wiliam's version is "The higher the stakes, the worse the assessment". There is evidence that this, while commonly true, is not inevitable.

[2] Ironically, what usually happens in practice is the reverse of parents choosing schools for the kids; because of limits of capacity in each school, popular schools choose their students.

[3] In health care it is now well-recognized that unbalanced targets distort clinical priorities. (For example, an earlier emphasis on reducing maximum waiting times had led to the treatment of some urgent cases being delayed.)

[4] If an English Language test were relabeled Mathematics, its "reliability" would be unchanged.
The statistical tools used measure consistency and levels of difficulty; they say nothing about *what is being assessed*.

[5] The education professions dislike the current tests so much that, seeking to marginalise testing, they make no serious effort to improve them.

of test-focused activities with regret; teachers regard it as inevitable – these are the measures of their performance which society has decided to value. Where balanced high-stakes tests have been adopted, they have proved a powerful influence in improving teaching and learning in every classroom (see Section 4).

The design challenge

The design of well-balanced assessment in a form that can be used for accountability purposes has been a solved problem for many years. There are working examples in the US and around the world of timed high-stakes examinations that show what can be done, and how it can enhance learning. They are not perfect but are vastly better balanced than most current tests. History contains many outstanding examinations that *enabled students to show what they know, understand and can do*[6]. The strategic design principle here is to include task types that represent the full range of performance goals.

The cost and complexity of high-quality balanced assessment is greater than for machine marked multiple choice tests; more complex tasks cannot be set and scored for $1 per student-test, a widely-accepted cost target in the US. (The massive cost of the class time wasted on otherwise-unproductive test-prep is usually ignored.)

There are well-established ways of lowering the cost of assessment so that it will monitor standards as reliably as at present, while *enhancing* student learning. A strategy that has multiple benefits is to make teachers the prime assessors, providing them with good assessment tools, and monitoring their scoring on a sampling basis. The many examples of this approach in practice show that it is also powerful professional development. It links naturally to formative assessment in the classroom which research shows to be such a powerful way of improving learning (Black and Wiliam 1998)

Strategically, it is unwise to hold costs for structured assessment down to current levels, well below 1% of the ~$10,000 per student-year that education typically costs. Feedback is crucial factor in determining the behaviour of systems of all kinds. Well–structured feedback on student achievement (Role A above), performance goals (Role B), and exemplar tasks for the classroom (Role C) is worth far more than the current investment in it.

Even when research-based methods of design and development have been used in assessment, notably in test development, the commissioning specification has often been too narrow, excluding design solutions that would allow the realization of the policy goals. The universality of the methods used in traditional psychometrics inevitably moves the focus from the kinds of performances that are assessed, which vary from subject to subject, to the statistical properties of the test.

---

[6] The Cockcroft Report (1982) defined good assessment in these terms.

How "standards" drive down standards

Many current models of national and state "standards" in mathematics and science are examples of bad strategic design – they have the effect in practice opposite to that intended. They actually drive down standards of performance in the subject. In explaining this I shall use as the lead example the National Curriculum in Mathematics in England. However, many state standards in mathematics in the US and elsewhere have much the same structure – and effect.

Criterion referencing is the source of the problem. The National Curriculum and most current "standards" in the US were designed on the principle that achievement goals can be specified through a detailed list of *level criteria* – concepts and skills that a student at that level should know, understand and show in tests. For example:

> Use the rules of indices for positive integer values, e.g.
> Simplify expressions such as $2x^2 + 3x^2$, $2x^2 \times 3x^2$, $(3x^2)^3$
> [1989 UK National Curriculum: in Algebra Target 2 Level 7]

or

> Factor simple quadratic expressions with integer coefficients, e.g., $x^2 + 6x + 9$, $x^2 + 2x - 3$, and $x^2 - 4$;
>
> solve simple quadratic equations, e.g., $x^2 = 16$ or $x^2 = 5$ (by taking square roots); $x^2 - x - 6 = 0$, $x^2 - 2x = 15$ (by factoring);
>
> verify solutions by evaluation. [Michigan Grade 8 standard A.FO.08.08]

Note the brevity of the task examples given.

Criterion referencing is an attractively simple idea. The public accepts it and policy makers on both sides of the Atlantic seem to love it[7]. But it is a dangerous illusion.

What is the problem? Fundamentally, it is that *the level of difficulty of a substantial task depends on various interacting factors – notably the complexity, unfamiliarity, and technical demand of the task, and the autonomy expected of the student in tackling it.* Thus the difficulty of the task is higher than that of its technical elements, tested separately – a complex task that is challenging for a good 16 year-old student (level 7) may require only mathematical concepts and skills that were taught in elementary school (level 4 and below). The "Consecutive Sums" task is an example[8].

Consecutive sums

The number 9 can be written as the sum of consecutive whole numbers in two ways:

9 = 2 + 3 + 4
9 = 4 + 5

The number 16 cannot be written as a consecutive sum.

Now look at other numbers and find out all you can about writing them as sums of consecutive whole numbers.

---

[7] When the National Curriculum was being developed, a senior UK policy maker was quoted saying: "Well, with maths, it's things you can either do or you can't, isn't it?" and went on to impose this checklist approach for Mathematics. Politicians and policy makers understand English Language much better – so essays and other extended writing are central, not vocabulary lists and grammar rules.

[8] Though more sophisticated concepts, such as the formalism for arithmetic progressions, can be used, most of the interesting results in this open problem can be found without them – and few 16-year–olds use them.

OK, but why are criterion-based standards *dangerous*? Because, it is only *fair* to give students the opportunity to meet the criteria for the highest level they might reach; this requires testing each concept and skill separately with short items. Thus the only way that "standards" which define levels for detailed concepts and skills can be made to work is by teaching and testing them separately in short items, like the following task [from Grade 10 GCSE]:

> (a) Factorise $x^2 - 10x + 21$
>
> (b) Hence solve $x^2 - 10x + 21 = 0$

Note the fragmentation of an already straightforward exercise, done to test explicitly two criteria:

(a) Can factorise a quadratic expression
(b) Can solve a quadratic equation

That kind of fragmented performance now dominates math tests and, because the stakes are high, dominates classroom learning activities (Ofsted 2006, 2008). Further evidence that such fragmentation is commonplace can be found by comparing test items with the standards, as above.

Such tasks are rarely found outside the mathematics classroom. It is not clear that success with such fragments has any value; it surely does not guarantee success with the more substantial chains of reasoning which doing and using mathematics involves. To be useful in solving substantial problems, from the real world or within mathematics, a technique needs multiple connections both to other math concepts and to diverse problem contexts, within and outside mathematics. These connections are built over time, by learning how to tackle more complex tasks like *Consecutive Sums*. Such task exemplars are much more challenging than their technical demand suggests because the strategic demand is a major part of the *total cognitive load* that determines difficulty.

In this way, a criterion-based approach drives down standards of overall mathematical performance. It undermines student learning by not preparing students to *think with mathematics* about the more substantial tasks they will meet in life outside the math classroom.

The design challenge

How might one design "standards" that set clear learning and performance goals without narrowing the curriculum? There have been various attempts at improving criteria to include strategic and tactical skills (often called *processes*) at different levels.

The extract shown is from the 2008 standards *(near below now %%)* of the Qualifications and Curriculum Authority (QCA) in England. Note the general descriptions of processes and the partial move away from detailed lists of techniques. It is clear that now any of the criteria can be interpreted at very different levels of difficulty. The tendency to narrow the task set remains – the easiest way to test, say, *representation* is separately, not as part of solving a substantial non-routine problem. Further, the processes do not change much across ages and levels – it is easy to find tasks that a typical 7 year old can do (~Level 2) that involve these processes – so the focus tends to remain on the content descriptions at each level.

Other countries have taken a quite different approach to the design of standards in mathematics and science, describing the learning and performance goals in broad terms. This approach relies on the professional expertise of teachers and

others to find a more detailed realisation that is appropriate to their local circumstances. The "flower diagram" used in the mathematics standards in Denmark illustrates this approach. These broad descriptions of competencies do not define levels of difficulty. So it is not surprising that they are common in school systems that do not use tests as an accountability tool with high-stakes consequences.

Since *difficulty is a property of the task*, *not its separate elements,* it can only be reliably determined by trialling the task with students, recognizing student responses at different levels in the scoring scheme. Any valid level scheme should be based on a set of well-analyzed tasks to which other tasks can then be related through trialling.

In an earlier paper *On specifying a curriculum* (Burkhardt 1990), prepared in the light of experience during the design of the National Curriculum, I pointed out that the final version gave no indication as to the types and balance of tasks that were to represent the performance goals in Mathematics – the concepts and skills could be shown entirely in short items, or in the course of three week long projects, or in a variety of other task types in between. I argued that to specify a curriculum relatively unambiguously, you need three[9] *independent* elements (see Figure 1):

- The tools in the toolkit of mathematical concepts and skills
- The performance targets, as exemplified by tasks
- The pattern of learning activities

They are independent, in that none of them determines the others, and complementary.

Currently in both the UK and the US there are attempts to produce improved models of standards. Of particular interest are the draft "College and Career Readiness Standards for Mathematics" developed for the Governors of US states as model national standards (%%link ref). This draft describes mathematical practices and principles in broad terms. Notably, it avoids detailed lists of technique, replacing them with a rich set of examples of tasks, covering a broad range of task types.

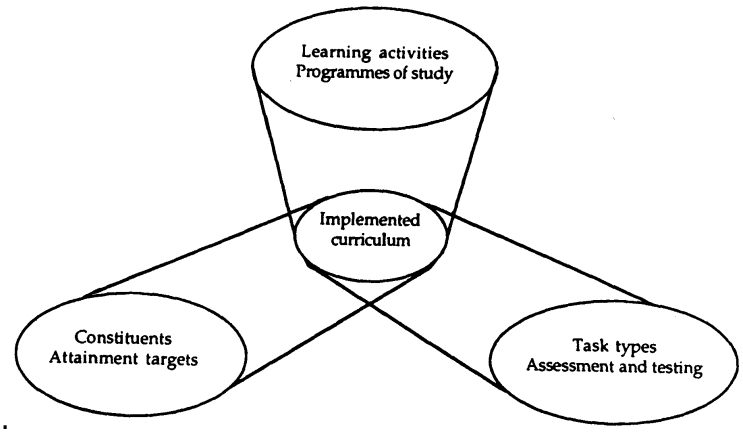

Figure 1. Three dimensions in specifying a curriculum

[9] 40 pages of varied task exemplars (typically 5 to 20 minutes) were included in the original version of the National Curriculum (DfES 1988), designed by the Government's Mathematics Working Group and circulated for comment. Their removal was never explained. The tests that emerged consisted entirely of short items, taking students an average of 90 seconds.