Section 3
What Does Assessment Assess?
Issues and Examples

It is said that a picture is worth a thousand words. When it comes to assessment, an example is worth that and more. Someone may claim to assess student understanding of X, but what does that mean? The meaning becomes clear when one sees what is actually being tested. Examples help for many reasons. They make authors' intentions clear, and they teach. In working through good examples of assessments, one learns how to think about student understanding. This section of the book offers a wide range of examples and much to think about.

In Chapter 6, Hugh Burkhardt takes readers on a tour of the assessment space. He asks a series of questions related to the creation and adoption of assessments, among them the following: Who is a particular assessment intended to inform? What purpose will it serve? (To monitor progress? To guide instruction? To aid or justify selection? To provide system accountability?) Which aspects of mathematical proficiency are valued? How often should assessment take place to achieve the desired goals? What will the consequences of assessment be, for students, teachers, schools, parents, politicians? What will it cost, and is the necessary amount an appropriate use of resources? Burkhardt lays out a set of design principles, and illustrates these principles with a broad range of challenging tasks. The tasks, in turn, represent the mathematical values Burkhardt considers central: specifically, that the processes of *mathematizing* and mathematical modeling are centrally important, as is the need for students to explain themselves clearly using mathematical language.

In Chapter 7, Jan de Lange continues the tour of mathematical assessments. Like Burkhardt, he believes that assessment development is an art form, and that like any art form, it follows certain principles, in the service of particular goals. He introduces the framework for the development of the Program for International Student Assessment. PISA assessments, like those of TIMSS (which formerly stood for Third International Mathematics and Science Study, and now for Trends in International...), are international assessments of mathematical competence. PISA differs substantially from TIMSS in that it focuses much more on students' ability to use mathematics in applied contexts. Hence, a design characteristic of PISA problems is that they must be realistic: the mathematics in the problems must correspond, in a meaningful way, to the phenomena they characterize. De Lange also argues that valuable assessments should highlight not only what students can do, but what they find difficult, sometimes pinpointing significant (and remediable) omissions in curricula.

In Chapter 8, Bernard Madison makes a somewhat parallel argument regarding the need for sense-making in an increasingly quantified world. For the most part, he notes, students exposed to the traditional U.S. curriculum have the formal mathematical tools they need in order to make sense of problems in context; what they lack is experience in framing problems in ways that make sense. This is increasingly important, as consumers and voters are bombarded with graphs and data that support contradictory or pre-determined positions. Full participation in a democratic society will call for being able to sort through the symbols to the underlying assumptions, and to see if they really make sense.

One virtue of cross-national studies is that is they raise questions about fundamental assumptions. People tend to make assumptions about what is and is not possible on the basis of their experience in particular contexts, which are often regional or national. Cross-national comparisons can reveal that something thought to be impossible is not only possible, but has been achieved in another culture. What needs to be done here to achieve it? In Chapter 9, Richard Askey uses a range of mathematics assessments to take readers on a tour of the possible. Some of these assessments are cross-national; others, which play the same role, are historical. It is quite interesting, for example, to compare the mathematical skills required of California teachers in 1875, and 125 years later!

In Chapter 10, David Foster, Pendred Noyce, and Sara Spiegel point to yet another use of assessment the way in which the systematic examination of student work can lead to teachers' deeper understanding of mathematics, of the strengths and weaknesses of the curricula they are using, and of student thinking. Foster, Noyce, and Spiegel describe the work of the Silicon Valley Mathematics Initiative (SVMI), which orchestrates an annual mathematics assessment given to more than 70,000 students. SVMI uses the information gleaned from the student work to produce a document called *Tools for Teachers*, which is the basis of professional development workshops with teachers. As Chapter 10 shows, such attention to student thinking pays off.

Readers of a certain age may remember the warnings that accompanied trial runs of the National Emergency Broadcast System: This is a test. This is only a test! The chapters in this section show that, properly constructed and used, assessments are anything but "only" tests. They are reflections of our values, and vital sources of information about students, curricula, and educational systems.

Assessing Mathematical Proficiency
MSRI Publications
Volume **53**, 2007

Chapter 6
Mathematical Proficiency: What Is Important? How Can It Be Measured?

HUGH BURKHARDT

This chapter examines important aspects of mathematical performance, and illustrates how they may be measured by assessments of K–12 students, both by high-stakes external examinations and in the classroom. We address the following questions:

- *Who does assessment inform?* Students? Teachers? Employers? Universities? Governments?
- *What is assessment for?* To monitor progress? To guide instruction? To aid or justify selection? To provide system accountability?
- *What aspects of mathematical proficiency are important and should be assessed?* Quick calculation? The ability to use knowledge in a new situation? The ability to communicate precisely?
- *When should assessment occur to achieve these goals?* Daily? Monthly? Yearly? Once?
- *What will the consequences of assessment be?* For students? For teachers? For schools? For parents? For politicians?
- *What will it cost, and is the necessary amount an appropriate use of resources?*

There are, of course, multiple answers to each of these interrelated questions. Each collection of answers creates a collection of constraints whose satisfaction may require a mix of different kinds of assessment: summative tests, assessment embedded in the curriculum, and daily informal observation and feedback in the

Malcolm Swan and Rita Crust led the design of many tasks in this chapter. The tasks were developed and refined in classrooms in the U.K. and the U.S. by the Mathematics Assessment Resource Service team. I have been fortunate to work with them all.

classroom. Rather than discuss each type of assessment, this chapter describes principles that should guide the choice of a system of assessment tasks created with these questions in mind, particularly the third: *What aspects of mathematical proficiency are important and should be assessed?* Every assessment is based on a system of values, often implicit, where choices have to be made (see [NRC 2001], for example); here I seek to unpack relationships between aspects of mathematical proficiency and types of assessment tasks, 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 type of task, although each requires very different kinds of mathematical proficiency.

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 makes little difference 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 the belief that "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. Moreover, the assumptions underlying their procedures are less-obviously flawed for short items. Teachers dislike all tests and want to minimize the time spent on them as a distraction from "real teaching" — ignoring the huge amounts of time they now spend on test preparation that is not useful for learning to do mathematics. Parents think "objectively scored" multiple-choice tests are "fairer" than those scored by other methods, ignoring the values and biases associated with multiple-choice tasks. None of these groups seems to be aware that assessment may affect students' learning of, view of, and attitude to mathematics. This chapter describes tasks that assess aspects of mathematical proficiency that may be difficult or impossible to assess by multiple-choice tasks.

[1] Only in the U.S., particularly in mainstream K–12 education. Other countries use much more substantial tasks, reliably scored by people using carefully developed scoring schemes.

Assess valued aspects of mathematical proficiency, not just its separate components. Measuring the latter tells you little about the former — because, in most worthwhile performances, the whole is much more than the sum of the parts. Is a basketball player assessed only through "shooting baskets" from various parts of the court and dribbling and blocking exercises? Of course not — scouts and sports commentators watch the player in a game. Are pianists assessed only through listening to 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 music or basketball, and should by the same token be assessed holistically as well as analytically. When we don't assess in this way (which, for U.S. school mathematics, is much 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 do these principles imply for assessment in 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 cannot find x without being led through the task by (a) and (b), is this worthwhile mathematics? For the student who can do the task without the aid of (a) and (b), this already-simple problem is further trivialized by fragmentation. Compare the triangle task to the following task, modified from the Balanced Assessment for the Mathematics Curriculum Project *Middle Grades Assessment Package 1* [BAMC 1999, p. 40], for students of the same age:

Consecutive Addends

> Some numbers equal the sum of consecutive natural numbers:
>
> $$\begin{aligned} 5 &= 2+3 \\ 9 &= 4+5 \\ &= 2+3+4 \end{aligned}$$
>
> - Find out all you can about sums of consecutive addends.

This is an *open investigation* of a surprisingly rich pure mathematical microcosm, where students have to formulate questions as well as answer them. It is a truly an *open-ended* task, i.e., one where diverse (and incomplete) solutions are expected, and can be used and assessed at various grade levels. (Note the crucial difference between an open-ended task and a *constructed response*.)

Scaffolding can be added to give students easier access, and a well-engineered ramp of difficulty, as illustrated by the following version.

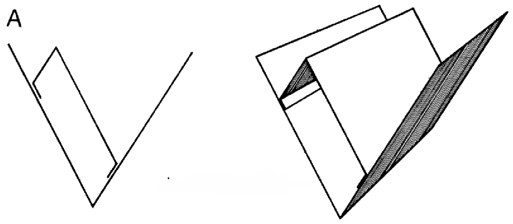
- Find a property of sums of two consecutive natural numbers.
- Find a property of sums of three consecutive natural numbers.
- Find a property of sums of n consecutive natural numbers.
- Which numbers are not a sum of consecutive addends?

In each case, explain why your results are true.

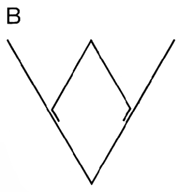
The proof in the last part is challenging for most people. However, the scaffolding means students only have to *answer* questions, not to *pose* them — an essential part of doing mathematics. Is this the kind of task 16-year-old students should be able to tackle effectively? What about the following task, modified from the *Be a Paper Engineer* module of [Swan et al. 1987–1989]? Is it worthwhile, and does it involve worthwhile mathematics?

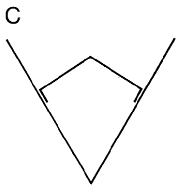
Will It Fold Flat?

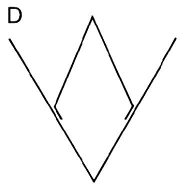
Diagram A is a side view of a pop-up card.

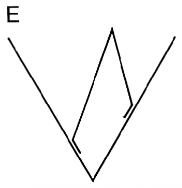


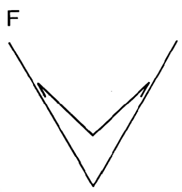
- Look at the diagrams below.
- Which cards can be closed without creasing in the wrong place?
- Which can be opened flat without tearing?
- Make up some rules for answering such questions.

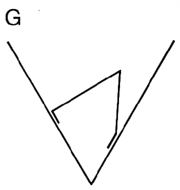


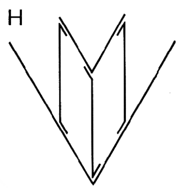


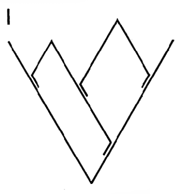








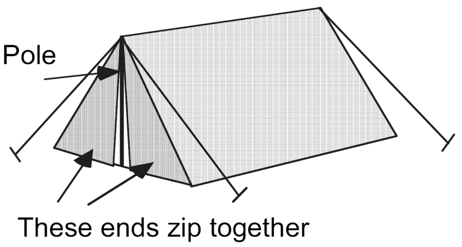




What about the following, more practical, task, adapted from the high-school level materials from the Balanced Assessment for the Mathematics Curriculum Project [BAMC 2000, p. 78]?

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). It should also be big enough so that someone can move around while kneeling down. Two vertical poles will hold the tent up.



Would the following more scaffolded version of the prompt in Design a Tent 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.) One might ask:

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

These tasks (particularly Will It Fold Flat? and Design a Tent) 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 mathematics.) The choice of performance targets, illustrated by the exemplars above, is at the heart of determining the content and nature of the K–12 mathematics curriculum. All sectors of society have an interest in these choices; mathematicians and mathematics educators need their views, and their informed consent. This requires the kind of well-informed debate that remains rare — and, too often, is obfuscated by the emotional over-simplifications of partisans on both sides of 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 the meaning of "correlate well"), it is *not* a justification for narrow tests. Why? Because assessment plays *three* major roles:

- A. to measure performance — i.e. "to enable students to show what they know, understand and can do;"

but also, with assessment that has high stakes for students and teachers, *inevitably*

- B. to exemplify the performance goals. Assessment tasks communicate vividly to teachers, students and their parents what is valued by society.

Thus

- C. to drive classroom learning activities via the WYTIWYG principle: What You Test Is What You Get.

The roles played by assessment have implications 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 classrooms — 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 takes A, B, and C into account. The roles played by assessment suggest that a system of assessment tasks should be designed to have two properties:

- *Curriculum balance*, such that teachers who "teach to the test" are led to provide a rich and balanced curriculum covering *all* the learning and performance goals embodied by state, national, or international standards.
- *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 teaching and learning high-quality mathematics. This is well recognized in some countries, where assessment is used to actively encourage improvement. In the U.S., a start has been made. The Mathematics Assessment Resource Service [Crust 2001–2004; NSMRE 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 a few aspects of mathematical performance, and that drive teaching and learning in the wrong directions. This is despite the fact that only a tiny fraction of educational spending is allocated to assessment.

It follows from B and C that choosing the range of task types to use in an assessment system together with lists of mathematical content is a rather clear way to determine a curriculum. (Lists of mathematics content alone, while essential, do not answer many key questions about the aspects of mathematical performance that are valued, so do not specify the types and frequency of assessment tasks. For example: What should be the balance of short items, 15-minute tasks, or three-week projects?) This issue and its relationship with a more analytic approach, are discussed in a later section.

Some common myths about assessment are worth noting:

Myth 1: Tests are precision instruments. They are not, as test-producers' fine print usually makes clear. Testing and then retesting the same student on parallel forms, "equated" to the same standard, can produce significantly different scores. This is ignored by most test-buyers who know that measurement uncertainty is not politically palatable, when life-changing decisions are made on the basis of test scores. The drive for precision leads to narrow assessment objectives and simplistic tests. (This line of reasoning suggests that we should test by measuring each student's height, a measure which is well-correlated with mathematics performance for students from ages 5 to 18.)

Myth 2: Each test should cover all the important mathematics in a unit or grade. It does not and cannot, even when the range of mathematics is narrowed to short content-focussed items; testing 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 (or learned) X but it wasn't tested this time." (Such sampling is accepted as the inevitable norm in other subjects. History examinations, year-by-year, ask for essays on different aspects of the history curriculum; final examinations in literature or poetry courses do not necessarily expect students to write about every book or poem studied.)

Myth 3: "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.

Myth 4: Testing takes too much time. This is true if testing is a distraction from the curriculum. It need not be, if the assessment tasks are also good learning (i.e., curriculum) tasks. Feedback is important in every system; in a later section 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 mathematics 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[2]; 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. (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.)

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 surprisingly consistent answers, and their priorities are not well-served by the current mathematics curriculum. I have space to discuss just a few key aspects.

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. American economic prosperity is said to depend 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 approximately $10,000 required for a K–12 mathematics education to develop the skills of machines that can be purchased for between $5 and $200. *Thinkers* appear to have more fun than drones, which is important for motivation. So, how do we assess *thinkers*? We give them problems that make them *think*, strategically, tactically, and technically — as will many of the problems student 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. Mathematics and mathematics teaching are two admirable and important professions — but their practitioners 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 K–12 mathematics curriculum, but should the design priorities be theirs, or more outward-looking ones that reflect society’s goals? The large amount of curriculum time devoted

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

to mathematics arose historically because of its utility in the outside world.[3] That priority, which now implies that the mathematics curriculum must change, should continue to be respected.

Mathematical Literacy

Mathematical literacy is an increasing focus of attention, internationally (see, e.g., [PISA 2003]) and in the U.S. (see, e.g., [Steen 2002]). The Organisation for Economic Co-operation and Development Programme of International Student Assessment, seeks to assess mathematical literacy, complementing the mathematically inward-looking student assessments of the Third International Mathematics and Science Study (see de Lange's chapter for more discussion of the design of these tests). Various terms[4] are used for mathematical literacy. In the U.S. "quantitative literacy" is common; in the U.K., where the term "numeracy" was coined [Crowther 1959], it is now being called "functional mathematics" [UK 2004b]. Each of these terms has an inherent ambiguity. Is it literacy *about* or *in using* mathematics? Is it functionality *inside* or *with* mathematics? The latter is the focus:

> *Functional mathematics* is mathematics that most *nonspecialist adults 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 not functional mathematics for most people. (If you doubt this, ask nonspecialist adults, such as English teachers or administrators, when they last used some mathematics they first learned in secondary school.) Functional mathematics is distinct from the "specialized mathematics" important for various professions.

The current U.S. curriculum is justifiable as specialized mathematics for some professions. However, as a gatekeeper subject, which is a key part of everyone's education, should mathematics education not have a large component of functional mathematics that every educated adult will actually use?

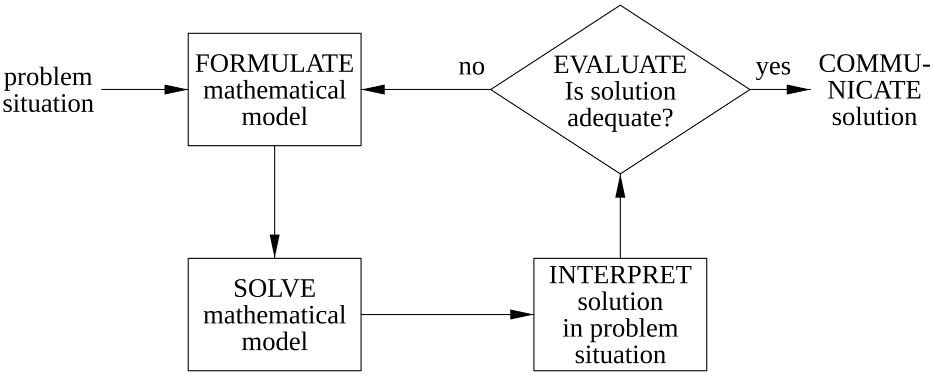
I shall outline what is needed to make the present U.S. high school mathematics functional, the core of which is the teaching of modeling. Modeling also reinforces the learning of mathematical concepts and skills (see [Burkhardt and Muller 2006]. This is not a zero-sum game.

[3] 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.

[4] 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.

Modeling

Skill in modeling is a key component in "doing mathematics." The figure below shows a standard outline of its key phases; see, for example, [Burkhardt 1981].[5] In current mathematics assessment and teaching, only the SOLVE phase gets much attention. (The situation is sometimes better in statistics curricula.)


problem situation
FORMULATE mathematical model
no
EVALUATE Is solution adequate?
yes
COMMUNICATE solution
SOLVE mathematical model
INTERPRET solution in problem situation

Key phases in modeling

Mathematical modeling is not an everyday term in school mathematics; indeed, it is often thought of as an 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 $5.00 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 the Eighth had 6 wives, how many wives did King Henry the Fourth have?

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 task, the answer depends on the type of oven (what remains constant: for traditional constant temperature ovens the answer is about 40 minutes, and for constant power microwave ovens there is rough proportionality, so an approximate answer would be $40 \div 5 = 8$ minutes).

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

For each problem, as usual, more refined models could also be discussed. For the third task, if students 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 modeled in mathematics classrooms should not involve specialist knowledge of another school subject but should be, as in the examples above, situations that children encounter or know about from everyday life. Teachers of English reap great benefits from making instruction relevant to students' lives; where mathematics teachers have done the same (see, e.g., [Swan et al. 1987–1989]), motivation is improved, particularly but not only with weaker students. Relationships in their classrooms are also transformed.[6] 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 Content Should We Include?

There will always be diverse views on content. This is not the place to enter into a detailed discussion of what mathematical topics should have what priority (for such a discussion see, [NRC 2001], for instance). Here I shall only discuss a few aspects of U.S. 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 time with mathematics, or should Euclidean geometry be considered specialized mathematics — an extra option for enthusiasts? Should not the algorithmic and functional aspects 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 mathematics and its many applications, or one for those whose future lies in the physical sciences and traditional engineering?

In the U.K., policy changes [UK 2004a; 2004b] have addressed such issues by introducing "double mathematics" from age 14, with a challenging functional "mathematics for life" course for all and additional specialized courses with a science and engineering, or business and information technology focus. It will be interesting to see how this develops. (The U.K. curriculum already has separate English language and English literature courses. All students take the first; about half take both.)

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

A 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*: open investigation; nonroutine problem; design; plan; evaluation and recommendation; review and critique; re-presentation of information; technical exercise; definition of concepts.
- *Nonroutineness*: 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.

Dimensions of Mathematical Performance

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, the Mathematics Assessment Resource Service has developed a *Framework for Balance*, summarized on the facing page. The *Framework* includes, as well as the familiar *content* dimension, the *phases of problem solving* from the figure on page 86, and various others including one holistic dimension, *task type*. This multidimensional analytic framework (it is dense, and takes time to absorb) is a way to examine how the major dimensions of performance are balanced in a particular test or array of assessment tasks. In most current tests, balance is sought only across the content dimension, and the only task type is short exercises that require only transforming and manipulating (the SOLVE phase).[7] The ability to formulate a problem is trivialized, and interpretation, critical evaluation and communication of results and reasoning are rarely assessed.

Task types. I will briefly illustrate the holistic dimension of the otherwise analytic *Framework for Balance* with tasks of each type. I chose to illustrate the holistic dimension because it brings out something of the variety of challenges that mathematics education and assessment should aim to sample (as in literature, science, social studies, music, etc.). Tasks are mostly given here in their core form rather in a form engineered for any specific grade. The tasks are designed 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 by including "open tasks" or "exponential ramps" to greater generality, complexity, and/or abstraction. We start the examples with two *planning tasks* — the second being 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 $100 per week to lease. Typical selling data is that one can sell an average of 30 ice creams per hour, each costing 50 cents to make and each selling for $1.50.

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

[7] 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 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.

Timing Traffic Lights

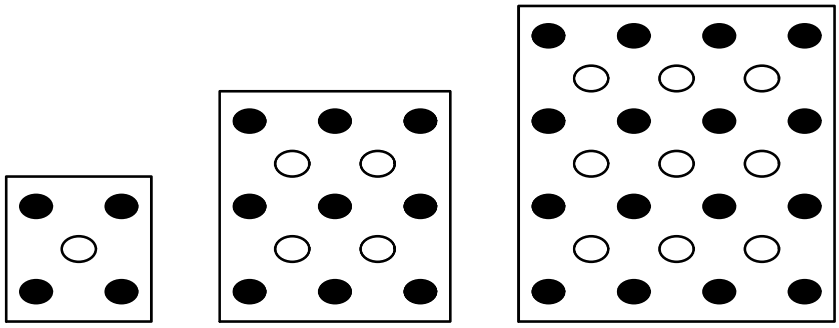
> A new set of traffic lights has been installed at an intersection formed by the crossing of two roads. Left turns are *not* permitted at this intersection.
>
> For how long should each road be shown the green light?

Treilibs et al. [1980] analyzed responses to these tasks from 120 very high-achieving grade 11 mathematics students and found that *none* used algebra for the modeling involved. (The students used numbers and graphs, more or less successfully.) These students all had five years of successful experience with algebra but, with no education in real problem solving, their algebra was non-functional. Modeling skill is important and, as many studies (see Swan et al. 1987–1989, for example) have shown, teachable.

The next task [Crust 2001–2004] is typical of a genre of *nonroutine problems* in pure mathematics, often based on pattern generalization, 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 ovals are dark chocolates and the white ovals are milk chocolates.



Box 1 Box 2 Box 3

> 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	13			

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

The scaffolding shown for this task fits the current range of performance in good middle school classrooms. One would hope that, as problem solving strategies and tactics become more central to the curriculum, part 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

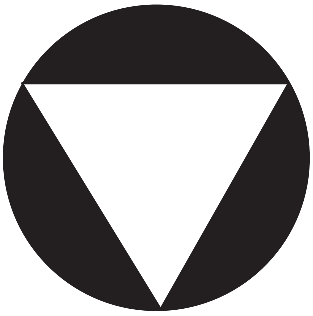
> Hans says "Olga has the longest average. She should go to the championship." Do you think Hans is right? Explain your reasoning.

This task provides great opportunities for discussing the merits and weaknesses of alternative measures. Ironically, in the TIMSS video lesson (from Germany, but it could be in the U.S.) on which this task is based, 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. (Bob Beamon — who barely qualified for the Olympics after two fouls in qualifying jumps — would have been excluded. He set a world record.) Is this good mathematics? I have found research mathematicians who defend it as "not wrong." What does this divorce from reality do for students' image of mathematics?

Magazine Cover [Crust 2001–2004] is a *re-presentation of information* task (for grade 3, but adults find it nontrivial). 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 for Magazine Cover illustrates how complex tasks can, with some scorer training, be *reliably* assessed — as is the practice in most countries and, in the United States, in some of the problems in the Advanced Placement exams.

Magazine Cover: Grade 3	Points
Core element of performance: describe a geometric pattern	
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
Allow 1 point for each feature up to a maximum of 6 points.	
Total possible points:	6

For our last example, we return to a type that, perhaps, best represents "doing" both mathematics and science — *investigation*. Consecutive Addends (page 79) 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 some of the things that people fear (or dream of), such as being

- struck by lightning
- murdered
- abducted by a stranger
- killed in a road accident
- a 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.

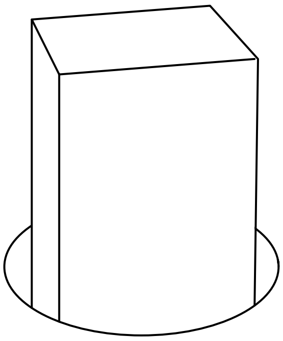
The tasks above, and the *Framework for Balance*, provide the basis for a response to our question, "What mathematics values should assessment reflect?" Taken together, they give a glimpse of the diversity of assessment tasks that enable students to show how well they can do mathematics — "making music"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

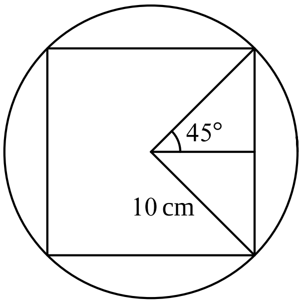
Square Peg

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 fit inside a circle of radius 10 cm.





- What percentage of the area of the circle is filled by the area of the square?

  Explain your work and show all your calculations.

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

Published examples of tasks of these various types include: a set of annual tests for grades 3 through 10 [Crust 2001–2004]; the *New Standards* exam-related tasks [NSMRE 1998], and classroom materials for assessment and teaching (*Balanced Assessment for the Mathematics Curriculum*: see, for instance, [BAMC 1999; 2000]). The *World Class Tests* [MARS 2002–2004] 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 can provide. Task design is usually subject to too-tight constraints of time and form. Starting with a good mathematics problem is necessary, but far from sufficient. As in all design: *Good design principles are not enough; the details matter*.[8]

[8] 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.

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 1986] contains some well-known benchmark examples.

The emphasis in this chapter on the task *exemplars* is no accident, but it is unconventional. However, 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, and the consensus is, “Yeah, our kids should be able to do that.”

Having looked in some depth at tasks that measure mathematical 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,[9] they need much better tools. For accountability, tests 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

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

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 assessment cost, and would this expenditure be an appropriate use of resources?*

Feedback is crucial for any complex interactive system. Systems that work well typically spend approximately 10% turnover on its "instrumentation." In U.S. education, total expenditure is approximately $10,000 per student-year, which suggests that approximately $1,000 per student-year should be spent on assessment across all subjects. Most of spending should be for assessment for learning in the classroom,[10] with about 10%, or approximately $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: budget shortage for administrators and dislike of assessment for teachers. Yet while "a dollar a student" remains the norm for mathematics assessment, students' education will be blighted by the influence of narrow tests. 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!

References

[BAMC 1999] *Balanced Assessment for the Mathematics Curriculum: Middle grades Package 1, Grade 6–9*, White Plains, NY: Dale Seymour Publications, 1999.

[BAMC 2000] *Balanced Assessment for the Mathematics Curriculum: High school Package 2, Grade 9–11*, White Plains, NY: Dale Seymour Publications, 2000.

[Burkhardt 1981] H. Burkhardt, *The real world and mathematics*, Glasgow: Blackie, 1981.

[Burkhardt and Muller 2006] H. Burkhardt and E. Muller, "Applications and modelling for mathematics", in *Applications and modelling in mathematics education*, edited by W. Blum et al., New ICMI Studies Series **10**, New York: Springer, 2006.

[Crowther 1959] Central Advisory Council for Education, *15–18: A report of the Central Advisory Council for Education* [Crowther Report], London: Her Majesty's Stationery Office, 1959.

[Crust 2001–2004] R. Crust and the Mathematics Assessment Resource Service Team, *Balanced assessment in mathematics* [annual tests for grades 3 through 10], Monterey, CA: CTB/McGraw-Hill, 2001–2004.

[10] 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."

[MARS 2002–2004] Mathematics Assessment Resource Service, *World class tests of problem solving in mathematics, science, and technology*, London: Nelson, 2002–2004. Shell Centre Team: D. Pead, M. Swan, R. Crust, J. Ridgway, & H. Burkhardt for the Qualifications and Curriculum Authority.

[NRC 2001] National Research Council (Mathematics Learning Study: Center for Education, Division of Behavioral and Social Sciences and Education), *Adding it up: Helping children learn mathematics*, edited by J. Kilpatrick et al., Washington, DC: National Academy Press, 2001.

[NSMRE 1998] *New standards mathematics reference examination*, San Antonio, TX: Harcourt Assessment, 1998.

[PISA 2003] Programme for International Student Assessment, *The PISA 2003 assessment framework: Mathematics, reading, science and problem solving knowledge and skills*, Paris: Organisation for Economic Co-operation and Development, 2003. Available at http://www.pisa.oecd.org/dataoecd/46/14/33694881.pdf. Retrieved 13 Jan 2007.

[Steen 2002] L. A. Steen (editor), *Mathematics and democracy: The case for quantitative literacy*, Washington, DC: National Council on Education and the Disciplines, 2002. Available at http://www.maa.org/ql/mathanddemocracy.html. Retrieved 28 Feb 2006.

[Swan 1986] M. Swan and the Shell Centre Team, *The language of functions and graphs*, Manchester, UK: Joint Matriculation Board, 1986. Reissued 2000, Nottingham, U.K: Shell Centre Publications.

[Swan et al. 1987–1989] M. Swan, J. Gillespie, B. Binns, H. Burkhardt, and the Shell Centre Team, *Numeracy through problem solving* [Curriculum modules], Nottingham, UK: Shell Centre Publications, 1987–1989. Reissued 2000, Harlow, UK: Longman.

[Treilibs et al. 1980] V. Treilibs, H. Burkhardt, and B. Low, *Formulation processes in mathematical modelling*, Nottingham: Shell Centre Publications, 1980.

[UK 2004a] Department for Education and Skills: Post-14 Mathematics Inquiry Steering Group, *Making mathematics count*, London: Her Majesty's Stationery Office, 2004. Available at http://www.mathsinquiry.org.uk/report/index.html. Retrieved 28 Feb 2006.

[UK 2004b] Department for Education and Skills: Working Group on 14–19 Reform, *14–19 curriculum and qualifications reform*, London: Her Majesty's Stationery Office, 2004. Available at http://publications.teachernet.gov.uk/eOrderingDownload/DfE-0976-2004.pdf. Retrieved 1 Mar 2006.