Proceedings

Conference on Mathematical Modeling

Teachers College
Columbia University

October 14, 2013

Editors:
Benjamin Dickman
Andrew Sanfratello

Table of Contents

Preface

Articles

Assessing Modeling

Hugh Burkhardt
MARS: Mathematics Assessment Resource Service
Shell Center, University of Nottingham and UC Berkeley

This paper reviews the challenges that higher-level skills present to assessment designers. The discussion is illustrated by exemplar tasks from two Shell Centre projects that emphasize modeling. From the various roles that assessment plays, the analysis starts with summative assessment, describing the multi-dimensional nature of task difficulty involving unfamiliarity, complexity and the degrees of autonomy expected of the student, as well as the technical demand. It distinguishes *expert*, *apprentice*, and *novice* tasks through the lengths of the chains of reasoning they involve and the expertise in mathematical practices they demand. An outline of the design principles for formative assessment that enhances learning includes some "do's" and "don'ts," leading to a look at future prospects and the strengths and limitations of computer-based assessment.

Keywords: Assessment; tasks; modeling; summative; formative; difficulty; computer-based

Introduction and Background

First, it's a great pleasure to be here, taking part in this tribute to Henry—a long-time friend and co-conspirator in moving forward the learning of modeling skills in K–12 education. The structure of this talk is: first some background, then some task exemplars, something on the roles of assessment, on summative and formative assessment and, finally, future prospects. As ever, I will use examples to clarify what I mean. Language tends to be interpreted within the reader's experience; examples make this more difficult. You will find a considerable overlap with Alan's talk (see page 13) because, although he's an insight-focused researcher and I'm an educational engineer, we have somehow managed to work together constructively since 1982. That, too, was Henry's doing—he persuaded the International Program Committee for the Adelaide ICME that we should lead the Problem Solving Theme. The quote was "Hugh and Alan deserve each other." We have to acknowledge the justice of that—who else would tolerate either of us for long? Because it's not really possible to understand things in detail in talks, you'll find this one will go at my usual warp speed. I want to cover a lot of ground, giving you a kaleidoscopic series of impressions and hoping you may find something useful among them.

Now the background. Henry and I have quite a lot of history in common. We both came to math education as mathematician-modelers, Henry from a distinguished position at Bell Labs and me as a theoretical physicist. We were early explorers of teaching modeling. I taught my first modeling course in 1963. Of course, like Monsieur Jourdain in the Molière play who was surprised to discover he'd been speaking prose all his life, most people have always done some modeling—but to be aware of it and to talk about teaching it is something else. And, well, Henry and I have been doing that ever since—fifty years of modeling, with some real progress but an awful lot still to do until it becomes a way of thinking for everybody.

We at the Shell Center are educational engineers; our aim is to find ways to help people who want to teach math better. For this to be scalable, things must be reproducible, namely materials. Now we all know that materials on their own aren't enough, but good tools can enable people to do things better, whether in the classroom, in professional development, or in assessment.

Assessment Tasks for Modeling

The examples I shall use come mainly from Shell Center projects on modeling, along with the assessment they involved. I shall concentrate on two projects that are modeling focused. *Numeracy Through Problem Solving* (Shell Center, 1987–89) was perhaps our greatest modeling project. It is materials-directed modeling, based on three-week modules with real outcomes. There are 5 modules: *Design a Board Game*, *Produce a Quiz Show*,[1] *Plan a Trip*, *Be a Paper Engineer*,[2] and *Be a Shrewd Chooser* about consumer decision-making. In each module the students work in small groups through four stages.

- Exploring the domain: This is done mainly through examples we provide, all with faults3 that the students identify.
- Generating ideas: The brainstorming phase, where each group develops its ideas.
- Refining the plan: This thing is going to happen and be judged, so detailed and careful planning and carrying through is essential.
- Evaluating the outcome: The students evaluate the products of the various groups.

Their work is directed by a "Student Guide," which is designed to present open challenges but to ensure, through checklists later in the Guide, that no essential step is left out—for example, the parent permission letter in *Plan a Trip*, or testing the questions in *Quiz Show* for fairness. The modules had real outcomes. The trips actually took place out of school, and the kids evaluated them. The board games were made as well as designed and they were played by the class, which evaluated each of them. Similarly with the "TV shows," which were great—though, happening in real time, this was the most challenging module for teachers to manage.

The assessment was of two kinds. In the course of the modules, there are assessment tasks that enable the teacher to check that all the students really understand, and are involved in, what is going on in their group. Figure 1 is an example from stage one of *Design a Board Game*, where students are learning about game design by finding faults in games we provided. Two of you play against each other. Your counters are on Start. You flip a coin. If you get heads, you move one square; if you get tails, you move two squares. What's wrong with this design?

Notice that there are some easy faults to find and some subtle ones—an example of the "ramp" of challenge that we aim to design into assessment tasks, so that all students can show what they can do at their own level. What we were trying to do was to see their ability to reason logically, which is very important in game design.

In addition to this "basic level" assessment, we had exams on these projects at the end of the course. There were two exams, at "standard" and "extension" levels, set by the exam board. This had a great advantage. All the kids had worked through the modules, so you knew that they had the background necessary, and you could

Coin Chutes and Ladders

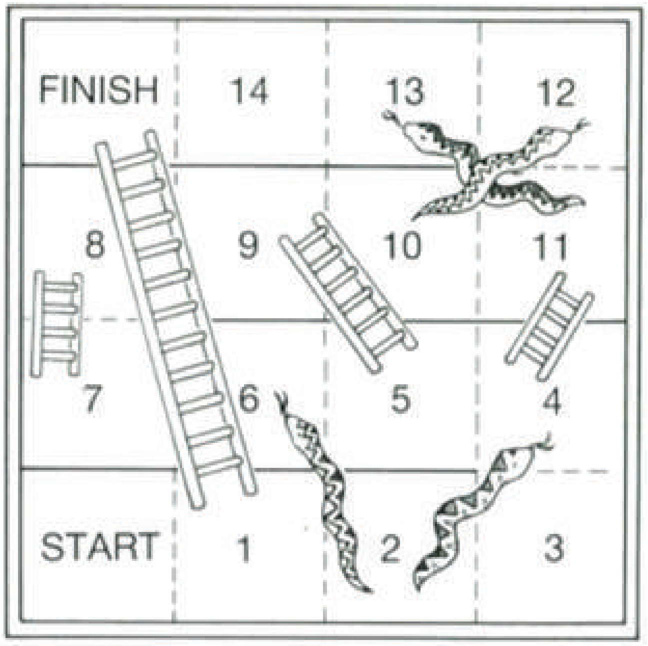


Figure 1.

[1] British for a TV game show, with contestants, questions and scores.

[2] *Paper Engineer* is about the design of pop-up cards and boxes. In it, students show geometrical reasoning at a much higher level than in an imitative geometry course.

[3] Faults are crucial. Commercial products, board games for example, are all highly refined so students can never match them. An unexpected spin-off was the students' delight at getting things with stupid errors from the examination board!

see how far they could generalize what they had learned—an example of assessing problem solving with a controlled "transfer distance."[4]

Figure 2 is an interesting voting systems problem from Plan a Trip. Choosing where to go for the class trip turned out, rather to our surprise, to be an important part of the work. Kids felt strongly and yet had to arrive at a class consensus on where to go, which demanded well-designed activities.

This task enables me to pay tribute to one of the great creations of modeling teaching and learning in the last fifty years, namely *For All Practical Purposes* from Sol Garfunkel and COMAP (I think, Henry was involved, too). If you don't know that book, then you certainly should. In it, there's an example, if I've got it right, of five candidates, five voters and five voting systems, each of which gives a different winner on the same data. That's clever.

Design a Tent (Figure 3) shows a rather old-fashioned tent. (This is a thirty-year old task.) I chose it to illustrate the challenge of designing a task that is open enough to be valid modeling, yet with sufficient guidance so that students do not go in too wildly different directions—important if you want to score responses reliably. Figure 3 gives the open version. If you set a problem as open as that, in what is supposed to be a fifteen-minute task in an exam, the kids get lost. So the version we actually used gave a little more scaffolding:

> Your task is to design this tent, drawing patterns for materials.
> The constraints are:
>
> - Big enough for two adults and baggage
> - Big enough to move around kneeling
> - Bottom a rectangle of plastic
> - Sides and the two ends will be made from a single sheet of canvas
> - Two vertical tent poles

The next example involves reasoning from data. It is based in a piece of software that, first, collects 100 reaction times from two "contestants," Joe and Maria (the teacher and a student works well), then enables the user to present the data in a variety of displays or summary statistics. The students' task is to use this same data to construct *two* arguments: (1) that Joe is quicker than Maria, and (2) that Maria is quicker than Joe.

This builds understanding of how data is used in politics and marketing—and thus intelligent scepticism about conclusions. This is an extraordinarily important insight in an era where our politicians, and even our businesses, try to avoid the "lie direct," but by careful selection and presentation of data, can make entirely opposite cases. The students also learn about the strengths and weaknesses of various ways of presenting data. Notably how summary measures (mean, median, mode, etc.) are far less informative than distributions of various kinds—throwing away most of the information, they are a relic from the days before computers. One distribution is particularly good for giving insight. You rank order each set of reaction times, then make a scatterplot of corresponding pairs. Differences show up in the position, slope, and curvature of the graph. [Officially *the percentile-percentile plot*, there must be better name; we call it the Gilbert diagram after the software designer's cat.]

My final example (Figure 4) in this brief *tour d'horizon* of assessment tasks that involve modeling is from our current Mathematics Assessment Project, which Alan has already introduced to you. Of course, this is a linear programming problem, but these students haven't been taught that technique. So for them it's a non-routine modeling problem (though it's already been "cleaned up" somewhat—you're given just the necessary information).

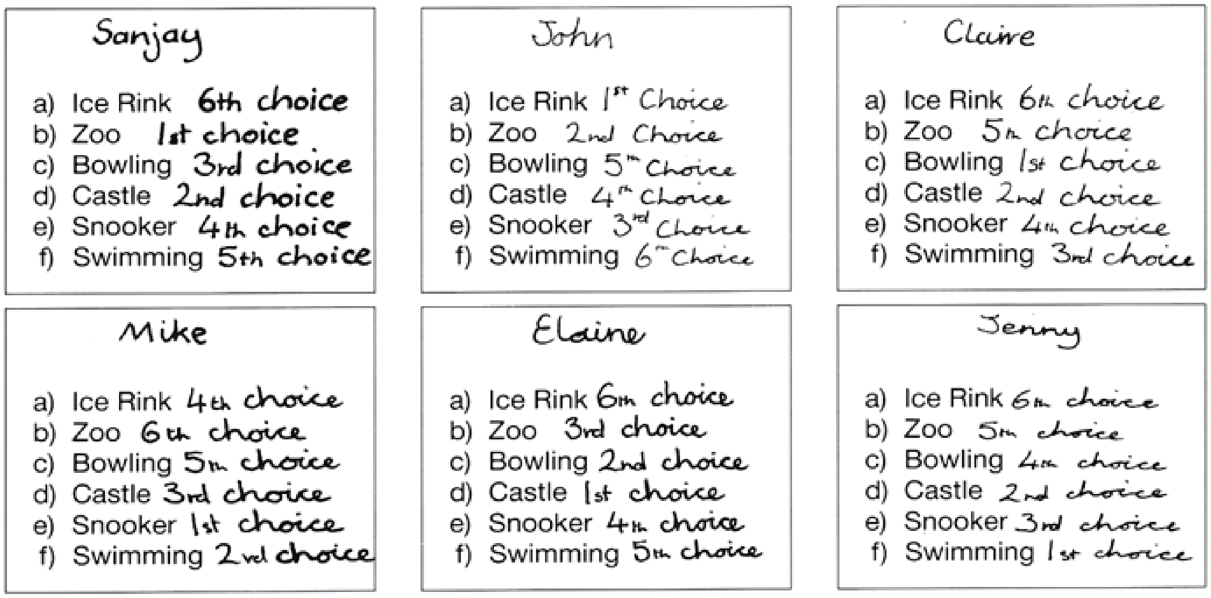
This leads me to an important distinction, shown in Figure 5.[5] Illustrative applications are, or should be, associated with every new mathematical topic—the focus is on the topic and you see various places it can be applied. That's very important for linking your mathematics to the real world, but it's not modeling; it's learning about models. In active modeling the situation is the focus. You have to decide which elements of your mathematical tool kit can be useful in tackling and understanding the situation, and then to find an effective way of using them in solving the problem. Autonomy of the solver is the essence.

[4] Transfer distance is a technical term representing how different two problems are.

[5] Thanks are due to Malcolm Swan, the Shell Center's lead designer and director, for this and so much else.

Plan a Trip: Voting

Six people are planning a day out.
Six different places have been suggested: Ice rink, Bowling alley, Swimming pool, Zoo, Castle, and Snooker hall.
They take a vote.
Which would be the best place for the trip and why?



Sanjay

a) Ice Rink 6th choice
b) Zoo 1st choice
c) Bowling 3rd choice
d) Castle 2nd choice
e) Snooker 4th choice
f) Swimming 5th choice

John

a) Ice Rink 1st Choice
b) Zoo 2nd Choice
c) Bowling 5th Choice
d) Castle 4th Choice
e) Snooker 3rd Choice
f) Swimming 6th Choice

Claire

a) Ice Rink 6th choice
b) Zoo 5th choice
c) Bowling 1st choice
d) Castle 2nd choice
e) Snooker 4th choice
f) Swimming 3rd choice

Mike

a) Ice Rink 4th choice
b) Zoo 6th choice
c) Bowling 5th choice
d) Castle 3rd choice
e) Snooker 1st choice
f) Swimming 2nd choice

Elaine

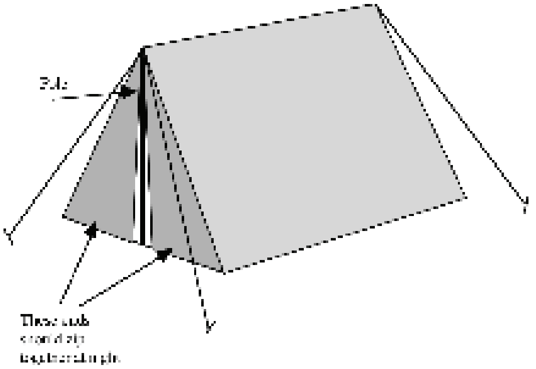
a) Ice Rink 6th choice
b) Zoo 3rd choice
c) Bowling 2nd choice
d) Castle 1st choice
e) Snooker 4th choice
f) Swimming 5th choice

Jenny

a) Ice Rink 6th choice
b) Zoo 5th choice
c) Bowling 4th choice
d) Castle 2nd choice
e) Snooker 3rd choice
f) Swimming 1st choice

Figure 2.

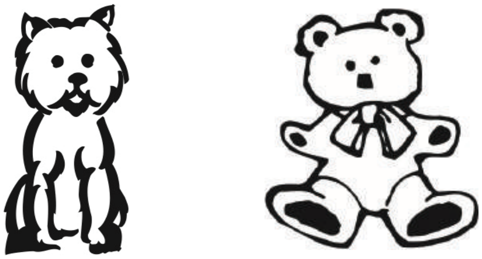
Design a Tent



Your task is to design this tent, drawing patterns for materials.

Figure 3.

Making Soft Toys



- Sue and Terry are making dogs and teddy bears.
- They have time to make 18 toys, and £60 to spend on materials
- Materials for a dog cost £3, materials for a teddy bear cost £4.
- They sell each dog for £8 and each teddy bear for £10.

How many of each should they make to maximise profit?

Figure 4.

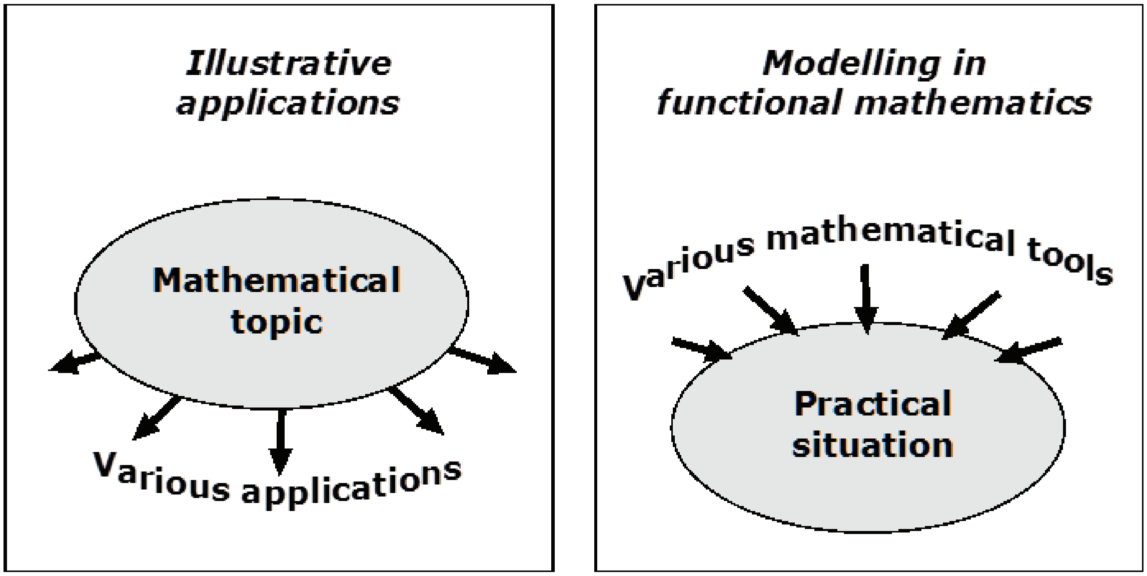

Illustrative applications
Mathematical topic
Various applications
Modelling in functional mathematics
Various mathematical tools
Practical situation

Figure 5.

Roles of Assessment

To get an overview of what I have covered so far, let's summarize the range and variety of modeling task types that could, and should, be included in the curriculum. They include (with examples) the following.

- Planning (Making Soft Toys)
- Design (Design a Tent)
- Evaluate and recommend (Reaction Times)
- Critique and improve (Coin Shoots and Ladders)
- Investigate (Voting methods)

I want to make a very important point that is too often ignored in assessment design: people can only autonomously use math they have thoroughly understood and connected—that does *not* normally include math they have just learned.[6] For students, there is a *few year gap* between the math they can do in an imitative exercise and the math they can call on when faced with a non-routine problem, whether in pure maths or modeling.

Now let us look at assessment more broadly. Of course, assessment tasks are much more than measurement tools. They are a compact way to show the performance goals of the curriculum. This is particularly true for high-stakes tests where, in most classrooms, *What You Test Is What You Get* (WYTIWYG). Teachers get criticised for "teaching to the test," but that is the main way our society judges them—their "bottom line."

Why do we assess? Well, there are various reasons, notably:

1. Summative assessment is *for reporting.* Its purposes are: to celebrate achievement; to reward effort and success; to select learners for groups, courses, or careers; to maintain records so that teachers, administrators, and parents can be informed of progress.
2. Formative assessment is *assessment for learning.* Its purposes are: to diagnose difficulties and so inform teaching; to motivate learners by showing them what they still need to learn. Formative assessment should be a week-by-week activity.[7]

[6] Unless you are a professional mathematician; they learn to use math they don't even know—but they know the method exists, and where to find it.

[7] Having frequent tests is better called *periodic assessment*; it is not formative assessment unless it leads to specific modifications of teaching in the light of what students show.

3. Evaluative assessment *for research.* Its purposes are: to assess teaching methods to see how they work in various circumstances; to gain further insight into teaching and learning; to develop tools and processes that enable teachers to work more effectively.

I'll say something about each of the first two, before going into each in more detail. Formative assessment needs rich detailed feedback. You want as much as you can, provided it's really informative, in a form that will motivate learners. The "give no scores" point surprised me very much, but the research is clear: If you give students scores, they don't pay heed to anything else. Interviews with kids said the same thing. For much summative assessment you only need one reliable number—for accountability purposes, nobody cares what it measures. This is clear because they've been giving "math tests" that, whatever they assess, don't assess mathematics. They assess some tiny fragments of mathematics that are easy to assess. Alan's told the ongoing story of that. One reliable number per student is plenty—often "they" only want one number per class (so they can sack the teacher) or per school (to justify "special measures").

Summative Assessment

The Common Core State Standards for Mathematics set out learning goals that represent both higher standards and greater challenges. Modeling is explicitly featured for the first time, and in two ways: as a *mathematical practice* at all grades and as a *content* label in high school. I like the authors' summary of what doing mathematics well involves:

> Proficient students expect mathematics to make sense. They take an active stance in solving mathematical problems. When faced with a non-routine problem, they have the courage to plunge in and try something, and they have the procedural and conceptual tools to carry through. They are experimenters and inventors, and can adapt known strategies to new problems. They think strategically. (Common Core State Standards Initiative, 2010, p. 4)

How far do our current tests assess this? Not far. So what should be done? The two assessment consortia, SBAC and PARCC, have built test specifications using a psychometric model based on "claims" and "evidence." Alan and I wrote the first draft "content specification" for Smarter Balanced (SBAC), the essentials of which have survived in the final version. Briefly, the four claims are:

- *Concepts & Procedures:* "Students can explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency."
- *Problem Solving:* "Students can solve a range of complex well-posed problems in pure and applied mathematics, making productive use of knowledge and problem solving strategies."
- *Communicating Reasoning:* "Students can clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others."
- *Modeling and Data Analysis:* "Students can analyze complex, real-world scenarios and can construct and use mathematical models to interpret and solve problems."

These are admirable performance goals (PARCC's are similar but less specific). Though they are well outside US assessment practice, good international examples of assessment from around the world show that, and how, they can be assessed in practice.

Since high stakes assessment is much more than "just measurement," exemplifying performance goals for most teachers, taking the standards seriously implies designing tests that meet them—"*tests worth teaching to*" that enable all students to show what they can do across the range of performances that the standards imply. This we call *balanced assessment*. It is defined through two strategic design principles.

- *Curriculum balance*: Design so that teachers who "teach to the test" will be led to deliver a curriculum that is balanced across learning goals.
- *Curriculum value*: Design test tasks that are valuable learning activities. People say, "Oh, we can't afford more than forty minutes. Learning time is so valuable." In most classrooms, the time they spend sitting for balanced examinations built from rich tasks will be one of the most valuable learning activities—tests should be designed so that is so.

Balanced Assessment for the Mathematics Curriculum was the first funded project that Alan and I, in 1992, got together for—and the work has been ongoing ever since in various forms.

Alan has already been nearly rude enough about psychometrics, but I would like to throw in my two pennies on the effect on test design. The standard psychometric view of "measurement theory" is too narrow, with an overwhelming focus on statistical error. In most scientific fields, two kinds of error are recognized: *systematic error* and *statistical error.* The accuracy of a measurement is determined by compounding these, not by choosing the smaller. In educational assessment, systematic error is not included. It consists in not assessing what you are really interested in, for example by focusing on what is easy to assess. Systematic error is not easy to quantify, so it is absorbed into the various uses of *validity.* The test of a test's value remains simple. In order of priority:

1. *Validity*: Is this a test worth teaching to?
2. *Reliability*: Can it provide one reasonably reliable score for each student?

But let's get back to design.

Summative Assessment—Design Challenges

Balanced assessment requires higher-level design skills than traditional testing—in everything but statistics. So how do you develop rich tests? First one must design tasks that involve good mathematics—thinking and reasoning and problem solving, using concepts and skills—along with rubrics that give appropriate credit. Then you observe students working on the tasks, talking with them, to find out the following.

- Did they understand the task as we intended?
- Were there unintended obstacles?
- Had we covered all reasonable approaches in the rubric?
- Did it enable students really to show what they can do?

Using this feedback, you revise the tasks and rubrics, iterating the process until the task works well. It's more like developing curriculum than short test items. (MAP 2012) gives many examples of rich tasks.

This brings me to another important fact that is often ignored in assessment design. The difficulty of a task depends on various factors. It makes a task more difficult if you increase the:

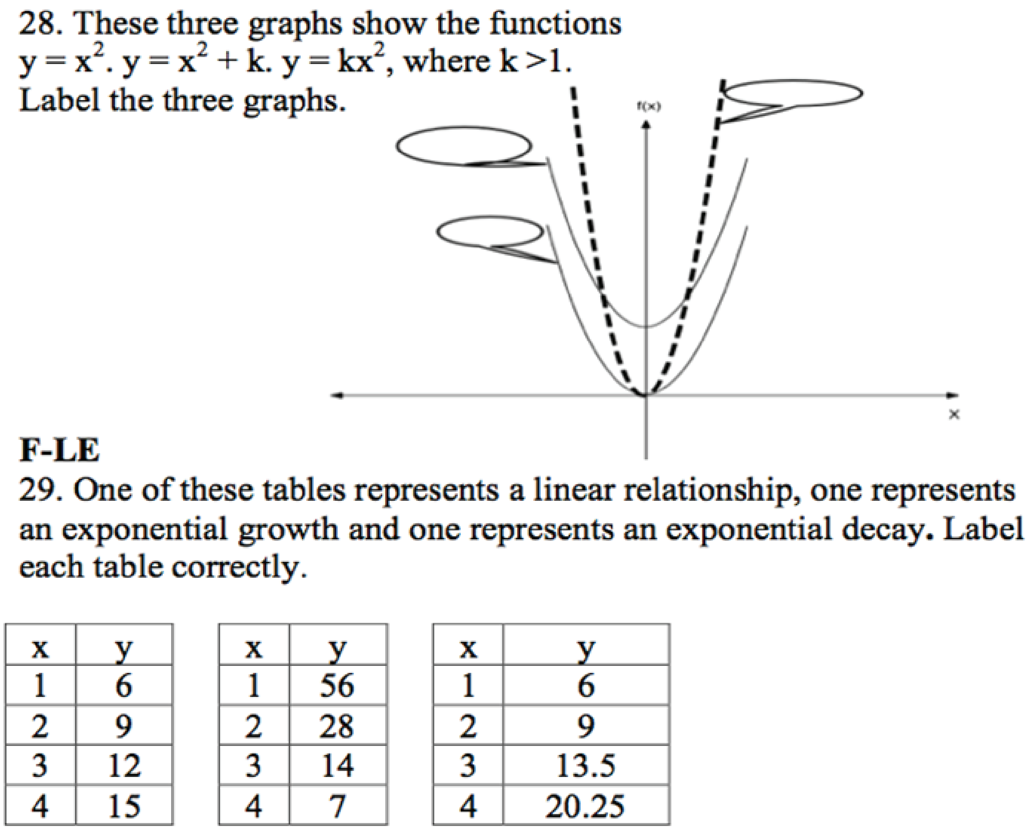
- *complexity*, by adding variables or more complex data, for example;
- *unfamiliarity*, a task that looks different from those you have solved before;
- *technical demand*, requiring higher-level math or longer chains of reasoning; and
- *autonomy*, what is expected of the student—with less guidance from the prompt or the teacher.

So while students may be expected to do a routine imitative exercise that tests recently learned "up to grade" math, they will not be able to use that math in a complex non-routine task that they are expected to solve without help. Again, this is because autonomous problem solving requires techniques to be thoroughly understood and *well-connected*—to standard applications and to other parts of mathematics. This takes a few years, even in classrooms that have moved beyond a "sequence of separate topics" approach to mathematics.

In light of all this, we have found it useful to distinguish tasks in a way that represents different emphasis on the mathematical practices.

- *Expert Tasks:* Rich tasks in a form they might naturally arise—in the real world or in pure mathematics (rather like writing extended passages in ELA). These are likely to fully involve the mathematical practices and all four aspects of difficulty, so they must not be too *technically* demanding. They assess the mathematics you will take out of the math classroom into the world beyond.
- *Apprentice Tasks:* Rich tasks but with scaffolding, structured so that students are guided through a "ramp" of increasing challenges. These involve the mathematical practices at a modest level, with less student autonomy.
- *Novice Tasks:* Short items, each focused on a specific concept or skill that has been learned (like spelling or grammar items in ELA). These present only technical demand, so this can be "up to grade," including concepts and skills just taught. They hardly involve the mathematical practices.

I chose these names because novices are learning the tricks of the trade. Experts solve problems as they arise. Apprentices do so under expert guidance. I'll illustrate these three types. The two novice tasks in Figure 6 are better



28. These three graphs show the functions $y = x^2$. $y = x^2 + k$. $y = kx^2$, where $k > 1$. Label the three graphs.

F-LE

29. One of these tables represents a linear relationship, one represents an exponential growth and one represents an exponential decay. Label each table correctly.

x	y
1	6
2	9
3	12
4	15

x	y
1	56
2	28
3	14
4	7

x	y
1	6
2	9
3	13.5
4	20.25

Figure 6.

than the average short item; they require a little thought—though elimination, which is not much used in really doing mathematics, remains the obvious strategy.

Short items are well understood. Now to where innovation is needed in US tests.

Expert Tasks

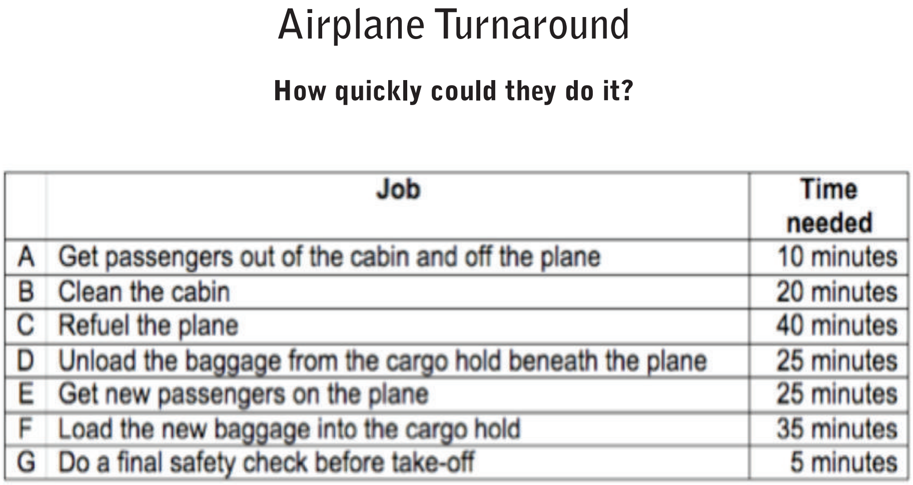
First, some expert task examples. *Airplane turnaround* (Figure 7) time gives the set of jobs they need to do on an airplane after it lands and before it takes off; how quickly could this be accomplished? To begin with, the students just add the times up. Then you give them a little bit more time. You ask, "Well, is that really the best you could do?" and some bright spark might say, "Well, while the passengers are getting out of the cabin and off the plane, could we unload the baggage?" And so on. Of course, this is a critical path analysis task in an informal problem solving form.

This task (Figure 8) involves

- formulating the problem mathematically,
- understanding exponential growth, and
- knowing it can't go on forever, and why.

Now you may be able to think of some fairly recent examples of people who have not understood that exponential growth can't go on forever. And some of them are in jail for a thousand years because of that. But anyway, it's a golden principle of exponential growth. The important thing is not just the growth rate. It's how it's going to crash. That's deep modeling understanding.

People talk about the difficulty of scoring complex tasks. Figure 9 is the scoring rubric for the Ponzi task. What we do with scoring rubrics is that we identify the core elements of performance. The total, by the way, is determined by how long the task should take. We find that a point per minute is about the right balance between giving points

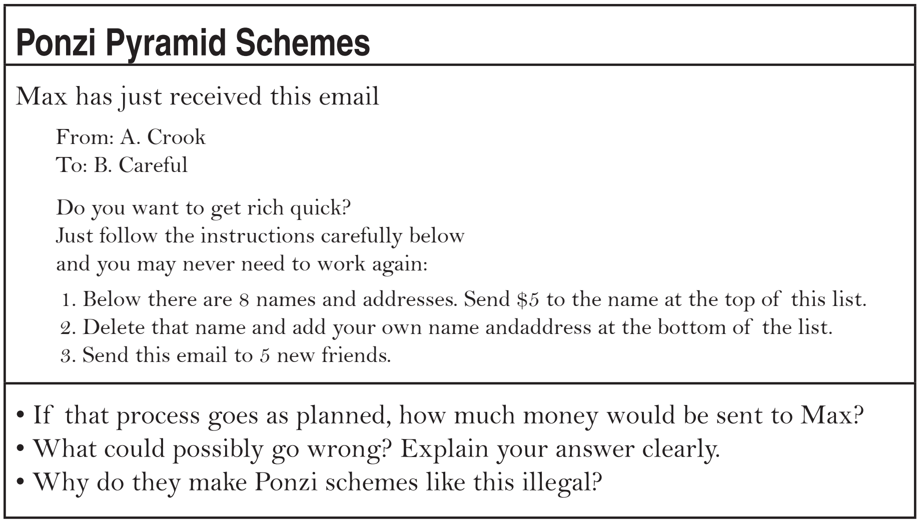


Airplane Turnaround

How quickly could they do it?

	Job	Time needed
A	Get passengers out of the cabin and off the plane	10 minutes
B	Clean the cabin	20 minutes
C	Refuel the plane	40 minutes
D	Unload the baggage from the cargo hold beneath the plane	25 minutes
E	Get new passengers on the plane	25 minutes
F	Load the new baggage into the cargo hold	35 minutes
G	Do a final safety check before take-off	5 minutes

Figure 7.



Ponzi Pyramid Schemes

Max has just received this email

From: A. Crook
To: B. Careful

Do you want to get rich quick?
Just follow the instructions carefully below
and you may never need to work again:

1. Below there are 8 names and addresses. Send $5 to the name at the top of this list.
2. Delete that name and add your own name andaddress at the bottom of the list.
3. Send this email to 5 new friends.

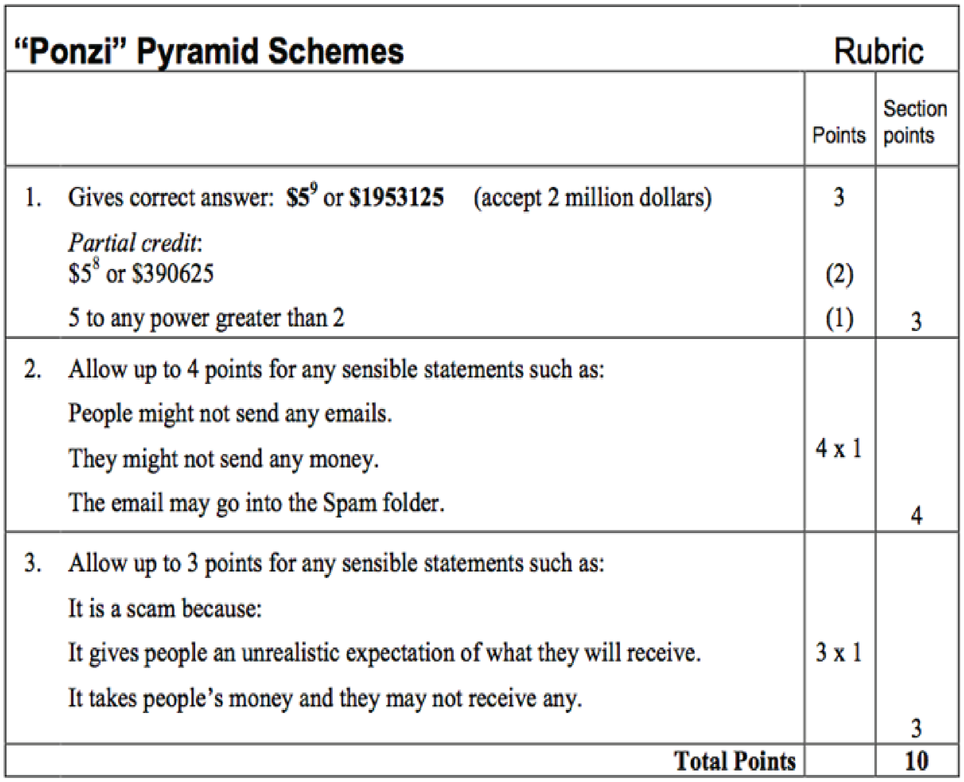
- If that process goes as planned, how much money would be sent to Max?
- What could possibly go wrong? Explain your answer clearly.
- Why do they make Ponzi schemes like this illegal?

Figure 8.

for different aspects of performance and having to make too-fine decisions. So part one gets three points, with partial credit if you get part of the way there. Notice seven of the ten points are not given for arithmetic but for modeling issues.

Proportional reasoning is at the heart of elementary modeling. The next task (Figure 10) looks like a very standard "three-number proportion" calculation. But the task for the students is different.

Asking students to critique other student responses is a design technique we find very powerful in various ways. Abdul, he's the good guy and explains his reasoning. The students' task is to explain the others' reasoning. Writing the units in helps. One of the great things in the Common Core State Standards is that Quantity, and thus continuous variables, is in from the very beginning, running parallel with counting variables: Number. I'll do Dorothy. Five over four; that was five dollars for four pounds, so that's dollars per pound. She uses something called the *unitary method*. How much for one pound? And then you multiply by the number of pounds. Now many teachers teach that as *the* method. In contrast, Stef first calculates a *scaling factor* from the weights, , then scales up the cost by that factor. All proportion problems are scaling problems—in some places they teach that approach, sometimes



“Ponzi” Pyramid Schemes	**Rubric**	
	Points	Section points
1. Gives correct answer: **$\$5^9$ or \$1953125** (accept 2 million dollars)	3	
Partial credit: $\$5^8$ or \$390625	(2)	
5 to any power greater than 2	(1)	3
2. Allow up to 4 points for any sensible statements such as:		
People might not send any emails.		
They might not send any money.	4 x 1	
The email may go into the Spam folder.		4
3. Allow up to 3 points for any sensible statements such as:		
It is a scam because:		
It gives people an unrealistic expectation of what they will receive.	3 x 1	
It takes people’s money and they may not receive any.		3
Total Points		**10**

Figure 9.

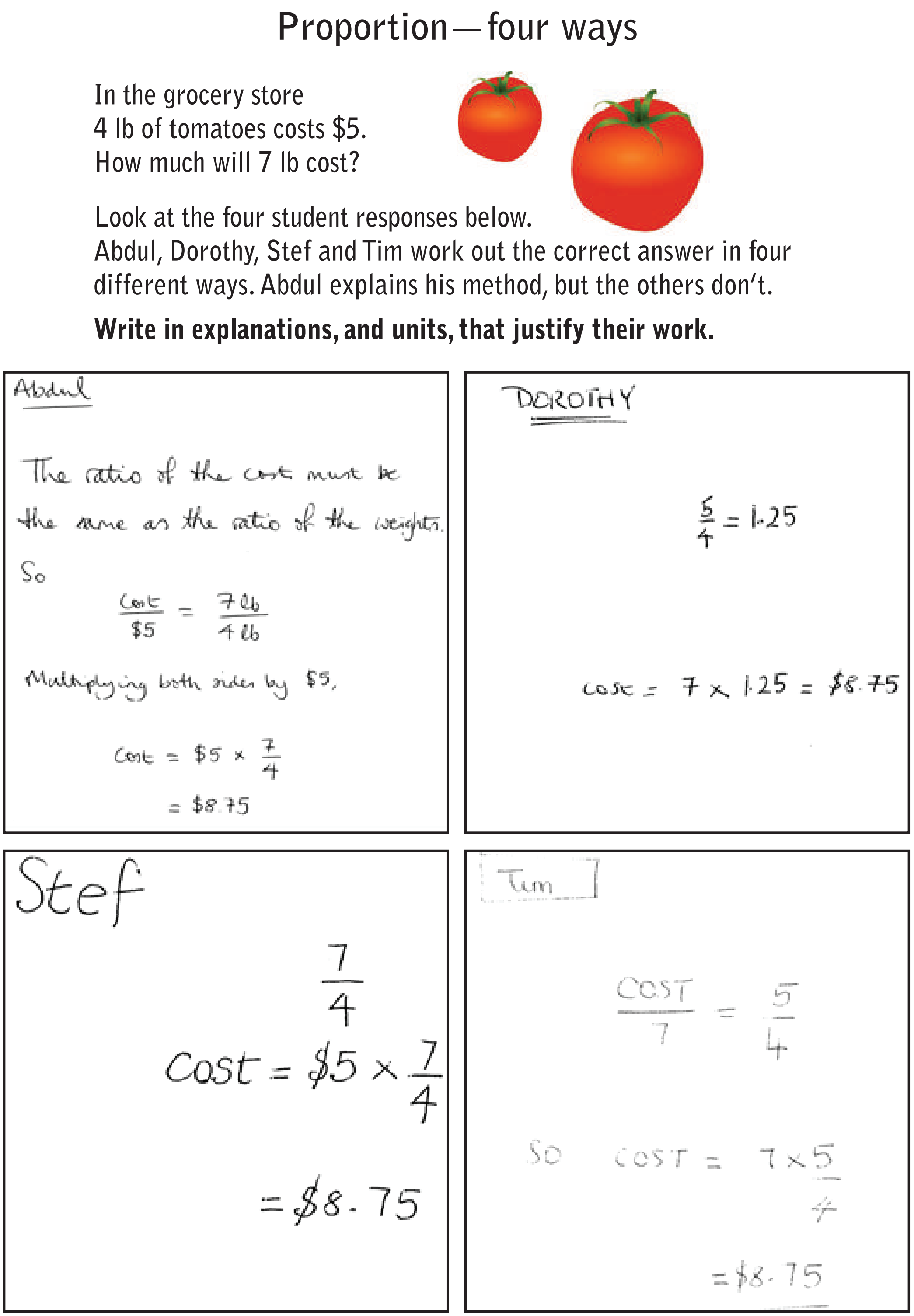
as a purely operational thing. The answer has one unit—dollars in this case. Of the three numbers you’re given, one has the same units as the answer, and the other two have the same units as each other. So it’s a scaling problem, and you just take the ratio of the last two and all you have to decide is whether it gets bigger or smaller. How about Tim? Put in the units and you’ll see that each side is the cost per pound expressed as ratios. Tim says (or rather doesn’t, but he should) that the *ratios will be equal*. Now I would say, and I’m not sure how well you score on this, that you only understand this form of proportional reasoning if you understand all those methods of thinking about it, that they will give the same answer, and why. Of course, the reason why they’re equivalent is that they are just rearrangements of the same mathematics.

One of the reasons why students find proportional reasoning harder than they should is that the modeling aspect is suppressed. Consider these three tasks.

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

So, fifty cents, okay, that’s “two-number proportion”; everybody can do that. Now if it takes forty minutes to bake five potatoes in the oven, how long will it take to bake one potato? Certainly, there are a number of students who will divide 40 five by 5. Is that wrong? Yes—unless it’s a microwave oven. It’s a modeling problem. And my favorite example: If King Henry VIII had six wives, how many wives had King Henry IV? (Laughs). Sometimes, there is no model—just facts, 2 (not 3) in this case, and with less drama.

But seriously, in the proportion units in every math textbook I know, all the problems will be proportion problems, which completely eliminates the modeling challenge that is otherwise clear in tasks like *Traffic Jam* (Figure 11). Notice that the rubric in Figure 12 again gives credit for the main modeling demands. Finally, for amusement, Figure 13: which sport might be represented by that speed-time graph? (another Malcolm Swan task, from 1984).



Proportion—four ways

In the grocery store
4 lb of tomatoes costs $5.
How much will 7 lb cost?

Look at the four student responses below.
Abdul, Dorothy, Stef and Tim work out the correct answer in four different ways. Abdul explains his method, but the others don't.

Write in explanations, and units, that justify their work.

Abdul

The ratio of the cost must be the same as the ratio of the weights.

So

$$\frac{\text{Cost}}{\$5} = \frac{7\text{ lb}}{4\text{ lb}}$$

Multiplying both sides by $5,

$$\text{Cost} = \$5 \times \frac{7}{4}$$

$$= \$8.75$$

DOROTHY

$$\frac{5}{4} = 1.25$$

$$\text{cost} = 7 \times 1.25 = \$8.75$$

Stef

$$\frac{7}{4}$$

$$\text{Cost} = \$5 \times \frac{7}{4}$$

$$= \$8.75$$

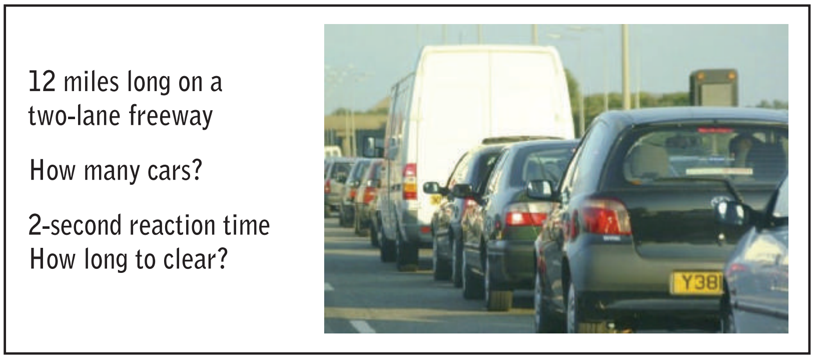
Tim

$$\frac{\text{COST}}{7} = \frac{5}{4}$$

$$\text{SO} \quad \text{COST} = 7 \times \frac{5}{4}$$

$$= \$8.75$$

Figure 10.



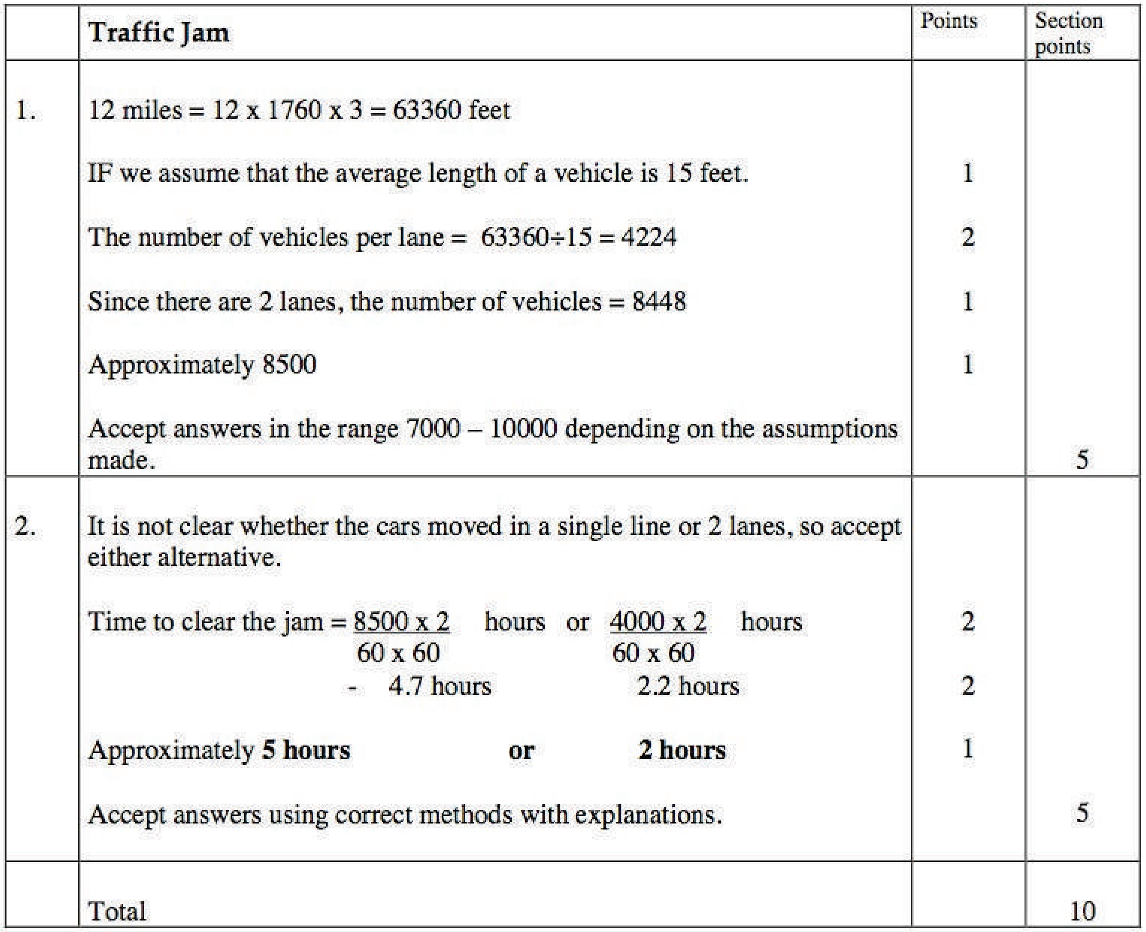
Traffic Jam

12 miles long on a
two-lane freeway

How many cars?

2-second reaction time
How long to clear?

Figure 11.



	Traffic Jam	Points	Section points
1.	12 miles = 12 x 1760 x 3 = 63360 feet		
	IF we assume that the average length of a vehicle is 15 feet.	1	
	The number of vehicles per lane = 63360÷15 = 4224	2	
	Since there are 2 lanes, the number of vehicles = 8448	1	
	Approximately 8500	1	
	Accept answers in the range 7000 – 10000 depending on the assumptions made.		5
2.	It is not clear whether the cars moved in a single line or 2 lanes, so accept either alternative.		
	Time to clear the jam = $\frac{8500 \text{ x } 2}{60 \text{ x } 60}$ hours or $\frac{4000 \text{ x } 2}{60 \text{ x } 60}$ hours	2	
	- 4.7 hours 2.2 hours	2	
	Approximately **5 hours** **or** **2 hours**	1	
	Accept answers using correct methods with explanations.		5
	Total		10

Figure 12.

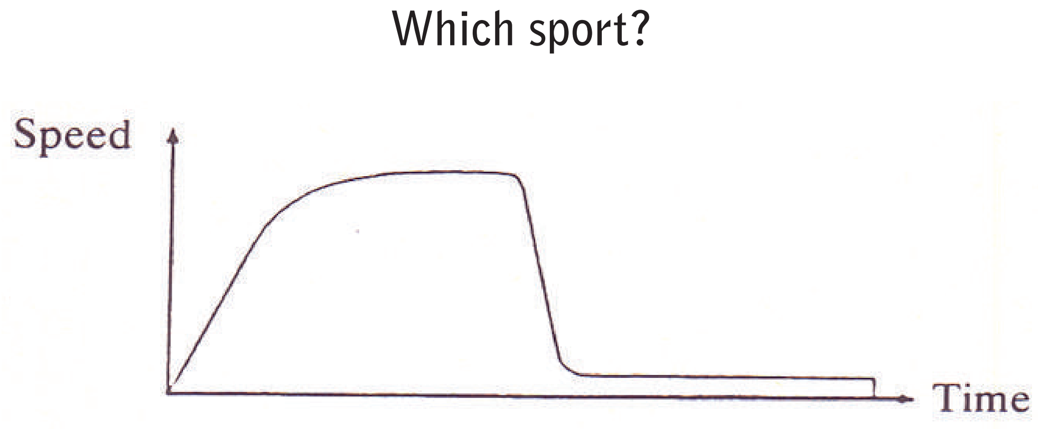


Figure 13.

Apprentice Tasks

Apprentice tasks are like climbing a mountain with a guide—an important transition to build expertise so you can climb without a guide. Since they take space, I'll show just one. The core task, finding and using a general rule, comes only at the end—the top of the ramp of challenge. The design of apprentice tasks involves guiding students through a ramp of challenge. "Patchwork" (Figures 14 and 15) does this by giving

- multiple examples that ease understanding;
- specific numerical cases to explore—by counting; and
- a helpful representation—the table.

Only then, does it

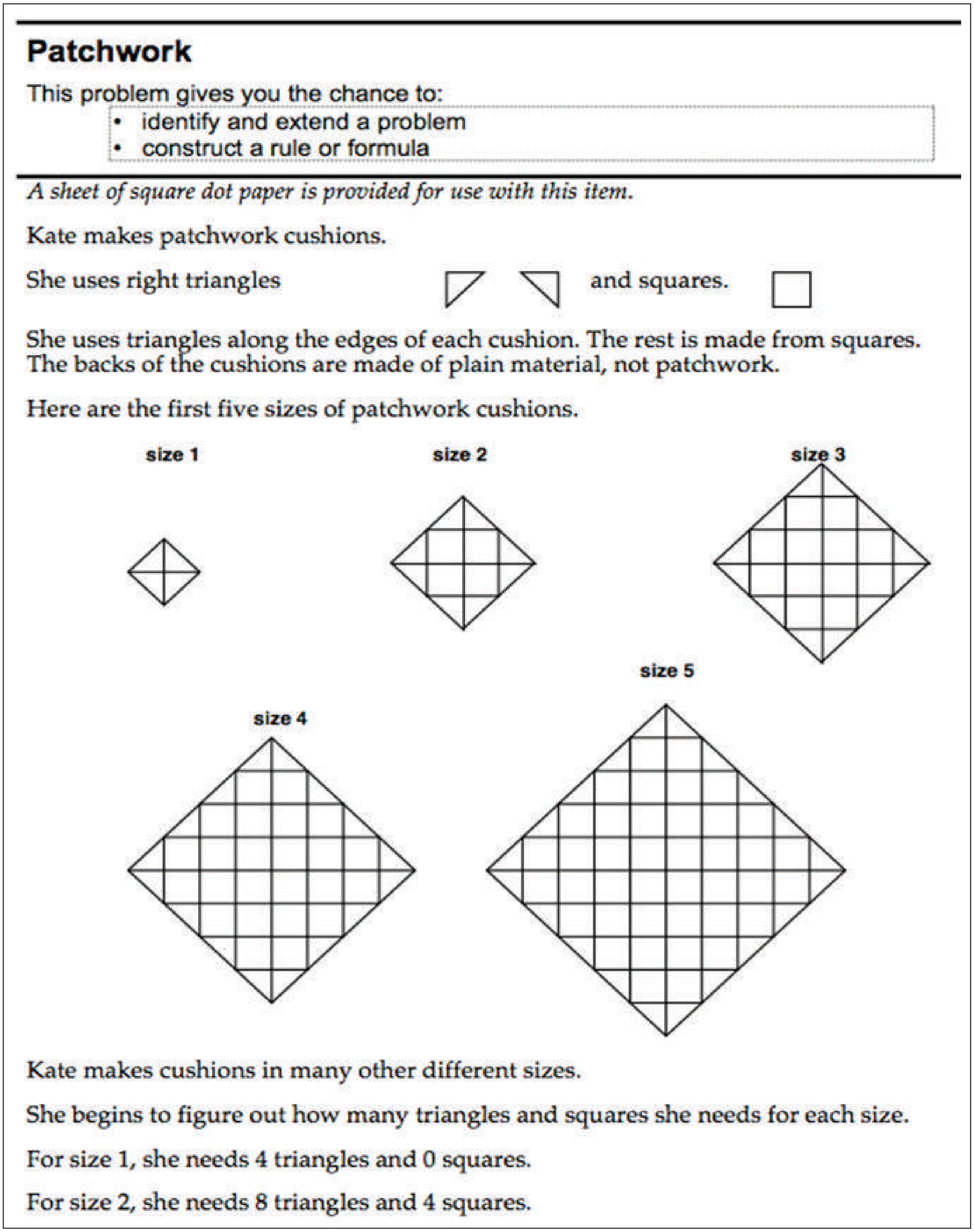
- ask for a generalization—rule, formula; and
- present an inverse problem.

Apprentice tasks assess growing expertise but, given the guidance, make only modest demands on the mathematical practices.

Computer-Based Testing

Computer-based testing is very much in fashion, promising cheap assessment under controlled conditions. Computers have great strengths and great weaknesses (see ISDDE 2012, Section 4). The following different phases of testing show both.

- Are computers good at the effective handling of the testing processes? They're brilliant.
- Do they provide better ways of presenting tasks? They can. You don't have to just use words, you could actually have animations and so on and so forth.
- Do they provide a natural medium for students to work on the math? No. If you have tried doing algebra on a computer, you know it's lousy. Diagrams too. (Things are getting better with tablets.)
- Are there effective ways of capturing a student response? Well, yes, if you scan the student's written paper in, but not if the computer wants to understand it.
- Are there reliable ways to score a student's response? Fine for novice tasks, but no good for complex tasks. When people say, "We've got computer scoring. It really works for complex tasks" be skeptical. People have been trying to do this for fifty years. I just send them a written student response with diagrams, notes and arrows and ask them to explain how they analyze it to give scores to the students. Never a reply. Cost effective human scoring works very well. Scan the stuff in and present it to scorers on screen with rubrics.
- On effective ways for collecting and reporting results, they're brilliant.



Patchwork

This problem gives you the chance to:

- identify and extend a problem
- construct a rule or formula

A sheet of square dot paper is provided for use with this item.

Kate makes patchwork cushions.

She uses right triangles ◸ ◹ and squares. □

She uses triangles along the edges of each cushion. The rest is made from squares. The backs of the cushions are made of plain material, not patchwork.

Here are the first five sizes of patchwork cushions.

size 1

size 2

size 3

size 4

size 5

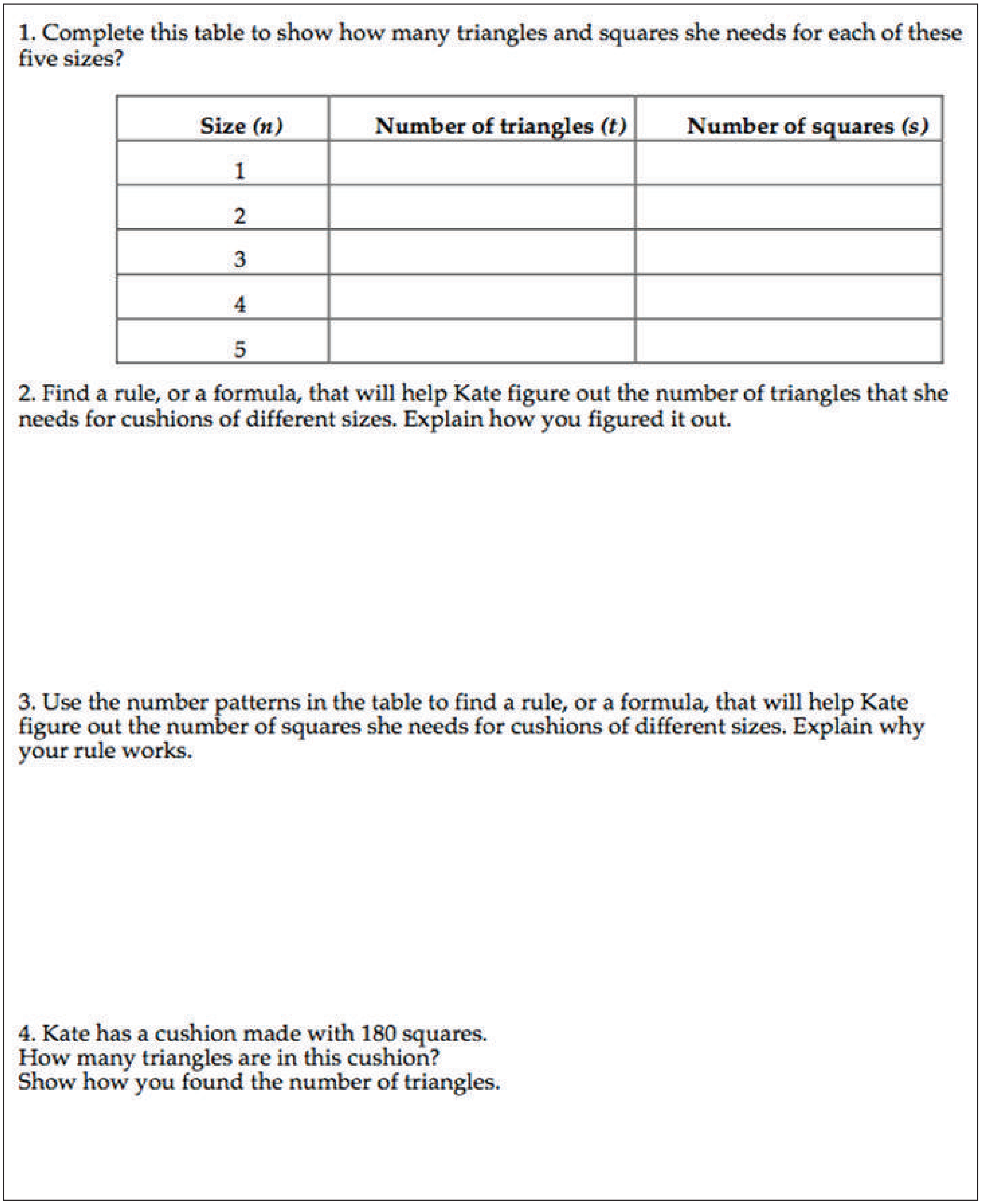
Kate makes cushions in many other different sizes.

She begins to figure out how many triangles and squares she needs for each size.

For size 1, she needs 4 triangles and 0 squares.

For size 2, she needs 8 triangles and 4 squares.

Figure 14.



1. Complete this table to show how many triangles and squares she needs for each of these five sizes?

Size (n)	Number of triangles (t)	Number of squares (s)
1		
2		
3		
4		
5		

2. Find a rule, or a formula, that will help Kate figure out the number of triangles that she needs for cushions of different sizes. Explain how you figured it out.

3. Use the number patterns in the table to find a rule, or a formula, that will help Kate figure out the number of squares she needs for cushions of different sizes. Explain why your rule works.

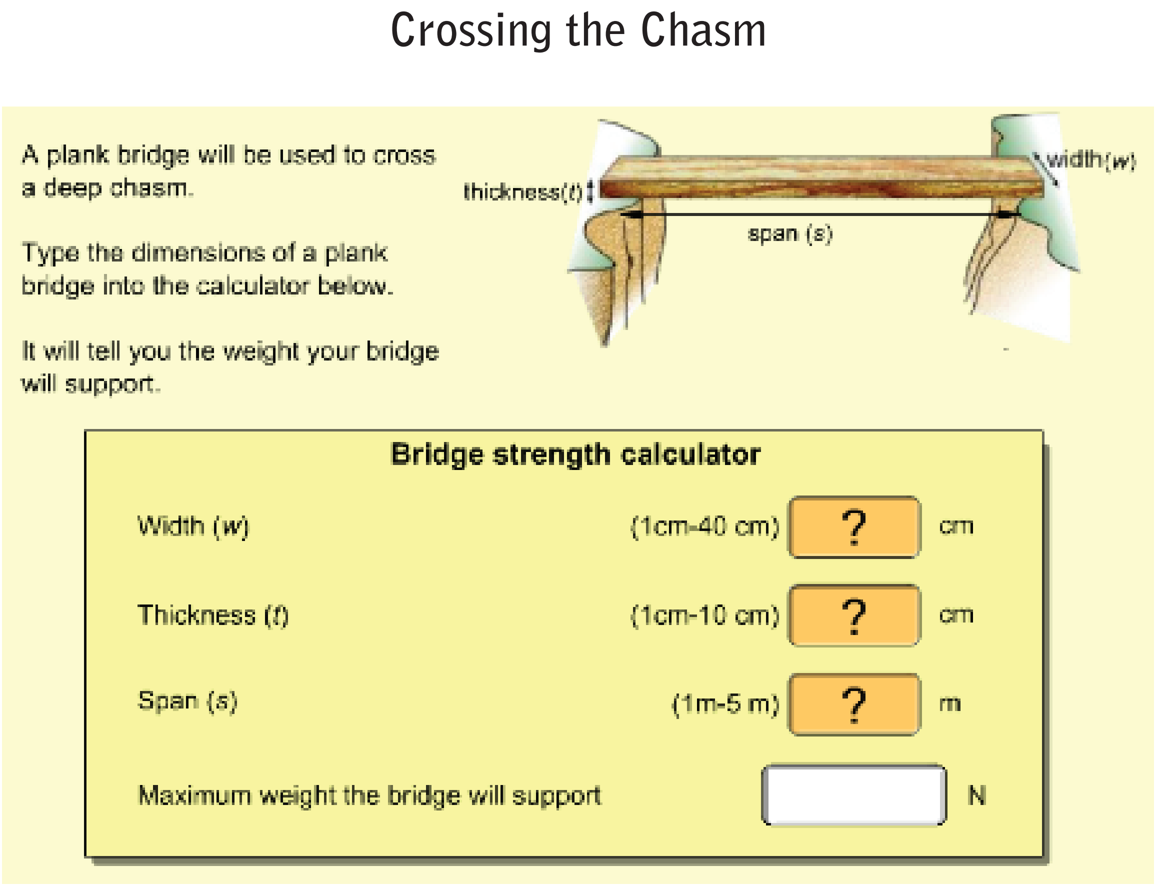
4. Kate has a cushion made with 180 squares.
How many triangles are in this cushion?
Show how you found the number of triangles.

Figure 15.

Crossing the chasm, a computer simulation, illustrates these points (Figure 16). The software enables students to explore how the bridge strength depends on each of the three variables. (There is a nice ramp of difficulty: the strength is proportional to the width—pretty obvious if you think of planks side by side; less obvious are *how* they increase with thickness and decrease with span, where students have to conjecture and check.) Figure 17 has examples of two different systematic approaches. It is important that the students had to both choose an appropriate representation and organize the results of their investigation in it. (Computer task designers would have to guide them, reducing the demand. Equally, scoring responses not in a standard form is beyond current or likely AI.)

Integrating Professional Development

There are cost issues, which I don't have time to say much about.[8] Let me just say that there is everything to be said for integrating professional development with assessment, and that scoring training can be particularly good professional development. The Hurdles Race (see page 17) task that Alan showed is one that we used a lot with teachers. We show the task, get the teachers to work on it, then show them six pieces of student work. We ask them to rank order them, then to construct a scoring rubric and score the responses. We then show them our scoring rubric. This leads to tremendous discussions about mathematics, pedagogy, and assessment. On that particular task, we threw in a wild card—we gave two points out of eight for the quality of the commentary. We had wonderful arguments about whether that deserves credit in a math test.



Crossing the Chasm

A plank bridge will be used to cross a deep chasm.

Type the dimensions of a plank bridge into the calculator below.

It will tell you the weight your bridge will support.

Bridge strength calculator

Width (*w*) (1cm-40 cm) ? cm

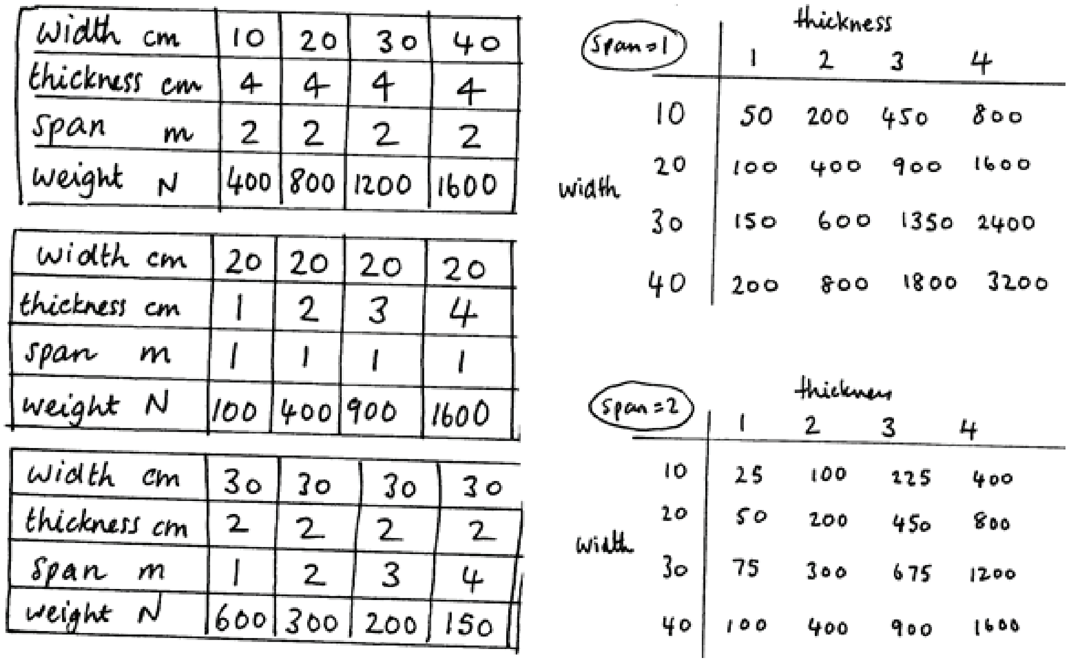
Thickness (*t*) (1cm-10 cm) ? cm

Span (*s*) (1m-5 m) ? m

Maximum weight the bridge will support N

Figure 16.

[8] For example, $1 tests are very expensive in that teachers use a lot of teaching time for otherwise unproductive test prep—with a real cost of around $200.



width cm	10	20	30	40
thickness cm	4	4	4	4
span m	2	2	2	2
weight N	400	800	1200	1600

width cm	20	20	20	20
thickness cm	1	2	3	4
span m	1	1	1	1
weight N	100	400	900	1600

width cm	30	30	30	30
thickness cm	2	2	2	2
span m	1	2	3	4
weight N	600	300	200	150

Span=1		thickness			
		1	2	3	4
width	10	50	200	450	800
	20	100	400	900	1600
	30	150	600	1350	2400
	40	200	800	1800	3200

Span=2		thickness			
		1	2	3	4
width	10	25	100	225	400
	20	50	200	450	800
	30	75	300	675	1200
	40	100	400	900	1600

Figure 17.

Formative Assessment for Modeling

So now for the second of the purposes I listed, formative assessment. Paul Black and Dylan Wiliam (1998, 2001) in their review of research—the thing that triggered all this activity—say that formative assessment is:

> . . . all those activities undertaken by teachers, and by their students in assessing themselves, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged. Such assessment becomes 'formative assessment' when the evidence is actually used to adapt the teaching work to meet the needs.

This research review shows that, *when well done*, formative assessment produces substantial gains in student learning. Now the challenge is this: most work to implement formative assessment has been done through professional development. Several years with skilled leaders were needed before "the penny began to drop for many of them." That's too expensive to do on a large scale, even if you could find the leaders. So our challenge in the Mathematics Assessment Project was to see how far materials could support formative assessment. The result is a set of formative assessment "lessons" that we call Classroom Challenges, twenty for each grade from 6 through high school.

Classroom challenges are of two kinds: *Concept development lessons* and *Problem solving lessons*, which include modeling. On the former I'll be brief. They are focused on sense making within mathematics. Why formative assessment? There's a lovely quote from one of the teachers in the trials:

> *I thought if I taught them all the pieces, they could put them together. I now know they can't.*

The missing component has been the *mathematical practices*, summarised in my earlier quote.

Problem Solving Lessons

Problem solving lessons have the following structure:

In 15–20 minutes of a prior lesson

- *each student tackles the initial, unscaffolded problem*—often a modeling problem. Students tackle the problem unaided. The work is collected in, not to be scored but to be looked through by the teacher so as to see what sorts of solutions are emerging.

In the main lesson

- *Pairs or groups work on the problem, comparing approaches.* Strategic hints are supplied to students that are still struggling.
- *Whole class discussion.* The payoff of mathematics is brought out by collecting and comparing solutions. This may be supported by students critiquing a range of sample responses, chosen by us to show a variety of approaches, including some using more powerful mathematics than is likely to have arisen in the class.
- *Individual work.* Building on the discussion, students improve their solutions to the initial problem, or one very much like it.

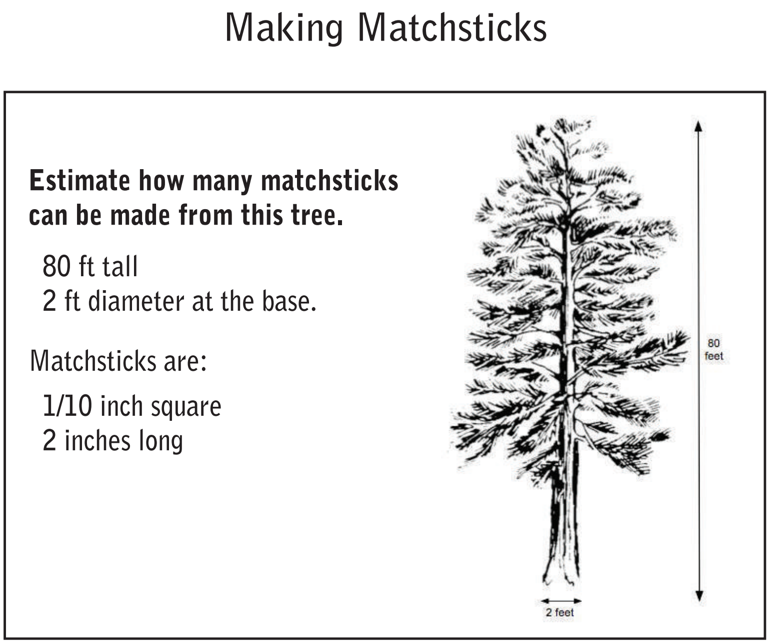
We aim to design materials that are open for the students but very supportive for the teachers, who are often outside their comfort zone in this kind of teaching. We suggest hints, because it's hard to invent good hints. They're always in the form of questions, not directions.

Teachers need both specific and generic support to achieve the changes we call "role shifting." Normally teachers play the *directive* roles: *manager, explainer,* and *task setter*, with students as imitative *responders*. The challenge of teaching problem solving and modeling is to move the teacher into *facilitative* roles: *counselor, fellow student,* and *resource*, with students becoming *investigators, managers, and explainers*. Piaget set out levels of understanding, roughly: *imitation, retention*—we're quite good on those—*explanation, adaptation,* and *extension*. We've not been good on the last three. Problem solving is about adaptation and extension, but explanation is crucial. The Common Core puts it well: the focus must move from answers to reasoning. Whenever you ask a student a question, a tremendous thing to do regularly is just to say, "Could you say that again?" or "Could you say a bit more?" You will often find a qualitative improvement—technically, the student turns an *exploratory explanation* into a *performance explanation*. They clarify what they're doing.

Let's look at an example, *Matchsticks* (Figure 18). We give them a formula sheet. There are a lot more formulae there than they need, but that makes them choose, giving them the option of treating the tree as a cylinder or a cone or whatever. At the discussion stage, we give them sample responses to discuss, asking each pair to consider:

- What has this student done correctly?
- What assumptions has (s)he made?
- How can (s)he improve the work?

When the students have done this problem, they enjoy engaging with others' reasoning. Alas, I don't have time to go through the responses but you will find it interesting to look at them on the website. One student, Jabir, basically misses on all sorts of dimensions, confusing feet and inches, volumes and lengths, not to mention the calculations—but he has done some things correctly. Sherida has done much better. She has a problem relating cubic inches to cubic feet. She also commits the, for me, unforgiveable sin of writing an answer that cannot be good to more than one significant figure with all the digits that her calculator showed. Appropriate accuracy is important, but rarely discussed. Chan is much more explicit about stating his assumptions. And so on.



Making Matchsticks

Estimate how many matchsticks can be made from this tree.

80 ft tall
2 ft diameter at the base.

Matchsticks are:

1/10 inch square
2 inches long

Figure 18.

Now what's the modeling process they're going through here? They have to:

- Understand the situation: that "fitting matchsticks into the tree" is a volume ratio problem;
- Choose a model for the tree: "Roughly a cylinder? Or a cone? Shall I ignore branches?";
- Calculate the volumes, using formulas;
- Get the units right, relating feet cubed to inches cubed;
- Compute the ratio, handling the big numbers—which students love but find hard;
- Choose appropriate accuracy—one significant figure;
- Evaluate their answer: "Does this make sense?"; and
- Ask: "How can we improve it? Do we need to?"

Where is the formative assessment? Teachers first see what students can do unaided—we help them analyze responses *without scoring them*. They then offer differentiated support to students as needed during the group work. Students get constructive feedback and, helped by the discussion of other students' work, they move their reasoning on. And so, finally, teachers get before and after comparisons.

I have to move on. Let me leave you with a lovely task, *Having Kittens*, and the samples of student work from the formative assessment lesson. Figure 19 is an advertisement outside a veterinarian's office. We talked about "multipliers" before (Laughs) and I asked what they were, so I thought I would show you the prize example (Laughs). Figure 20 has some samples of representations that the kids have tried to get on this exponential upon exponential problem. It generates wonderfully rich mathematical discussion. (Try it for yourself. You'll find the figure is realistic, even ignoring what I call the Randy Tom effect—the potentially enormous number that one male descendent and all the females in the neighborhood might produce!)

Summarizing, it's not a trivial matter to produce one of these formative assessment lessons, because formative assessment needs "adaptive expertise." Most teachers need both specific and generic support to develop the range of mathematical and pedagogical skills needed. Results are encouraging, with millions of lessons downloaded[9] by teachers and teacher leaders so far. This work has involved a design team of about eight people working at the Shell Center, led by Malcolm Swan. Each Classroom Challenge costs about twenty-five thousand dollars to develop. Why so expensive? The "authorship" approach works for a brilliant teacher in their own classroom where they can get the feedback as they go along. But if you want stuff to be both ambitious and robust in the hands of more typical teachers, you have to trial it through cycles of development like any other product—and that's expensive. Within MAP, there's a broader team with a research program that Alan has referred to. The project is directed by Alan and me, with Malcolm Swan, Daniel Pead, and Phil Daro—the chair of the standards writing group for the Common Core. It shows the sort of scale needed to turn ambitious goals into robust products.

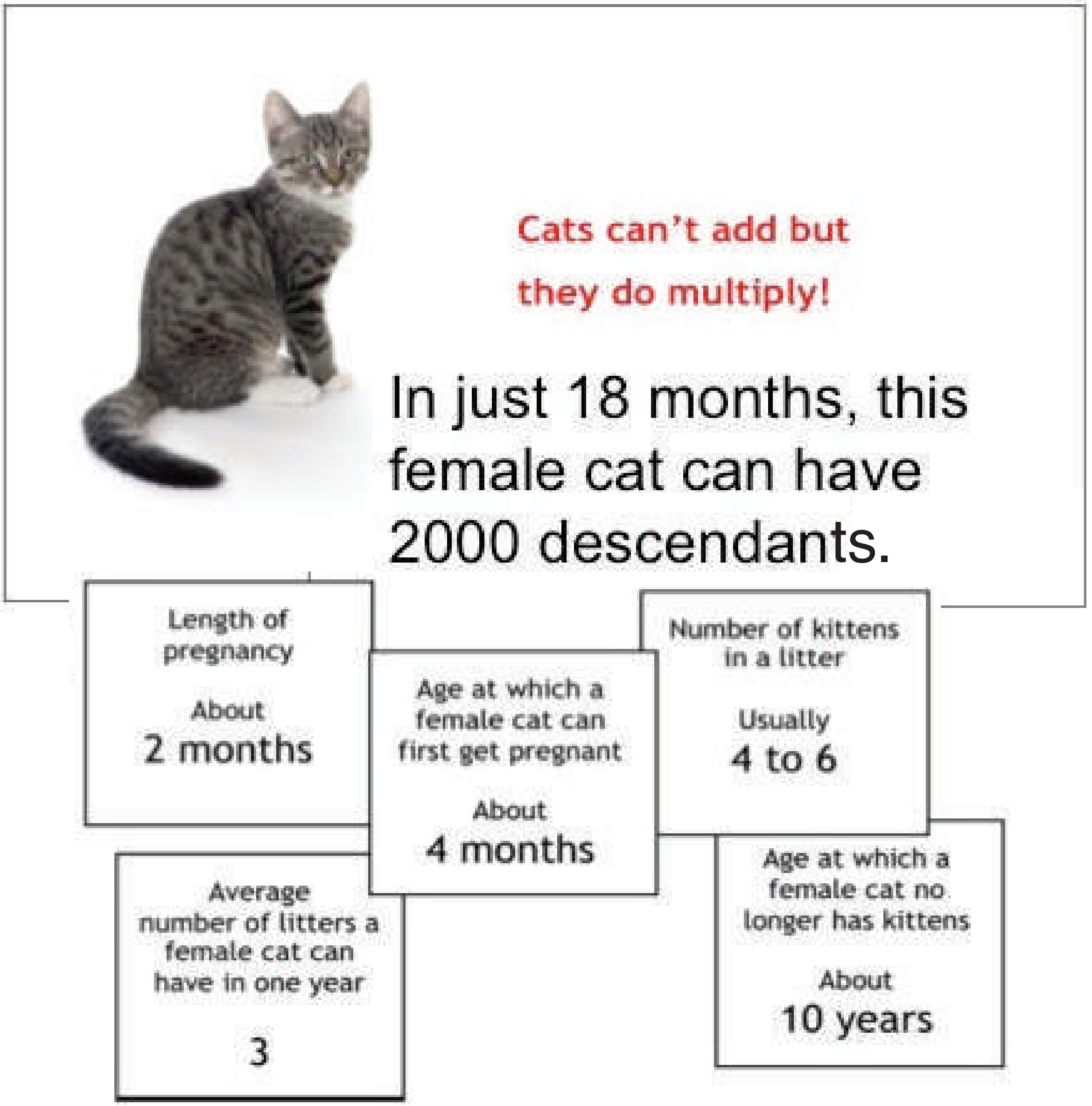
Future Prospects

Now to future prospects. As Alan just hinted, the signals are mixed. The learning goals embodied in the Common Core are admirable, as are the performance goals in the Smarter Balanced content specification. Realizing these goals will take many people outside their comfort zones, particularly in the design of rich tasks that will work well with kids in exams, or indeed in lessons, and in implementation that works smoothly.

The whole process must have public credibility. For example, SBAC initially chose to have a computer-adaptive test plus, for assessing higher-level skills, two days of classroom performance tasks, handled and scored by the class teacher. They then said that the classroom tasks weren't going to count in the final scores. But even hadn't they said that, the classroom assessment scores would not have credibility unless you took great care to engineer the circumstances to insure that the students were indeed doing the tasks themselves and not having Mom or Pop doing it, or the teacher feeding them the answers. So one of our victories, we think, is that there will be a timed, written examination as part of the SBAC tests, scored by humans, for the assessment of Claims 2, 3, and 4. One of the current dangers is SBAC overestimating the in-house expertise of testing companies. People do what they know how to do, so psychometricians will do Item Response Theory, and test designers will design short items because that's what they've always done. To develop rich performance tasks requires bringing in new skill sets, which needs

[9] Free for non-commercial use from (MAP, 2012)

Assessment Task: Having Kittens



Cats can't add but they do multiply!

In just 18 months, this female cat can have 2000 descendants.

Length of pregnancy

About 2 months

Age at which a female cat can first get pregnant

About 4 months

Number of kittens in a litter

Usually 4 to 6

Average number of litters a female cat can have in one year

3

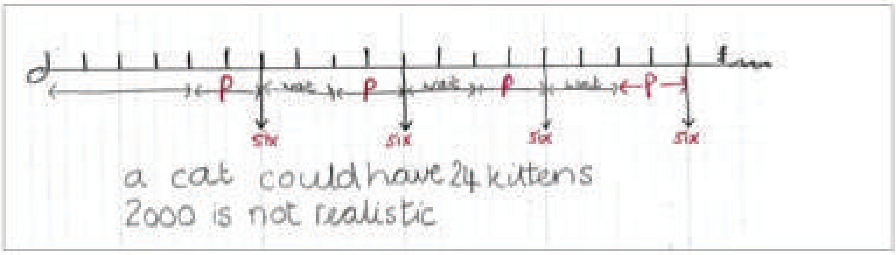
Age at which a female cat no longer has kittens

About 10 years

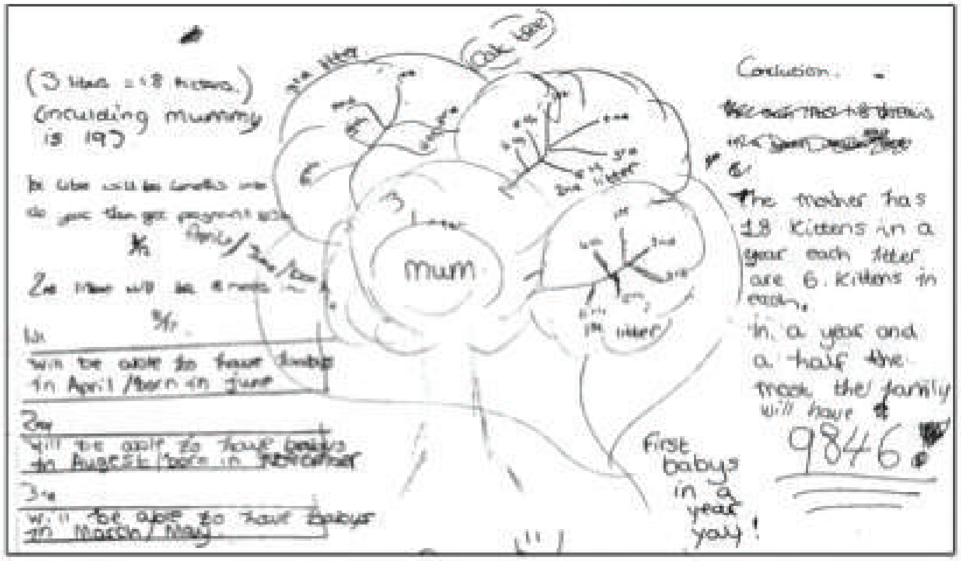
Is this figure of 2,000 realistic?

Figure 19.

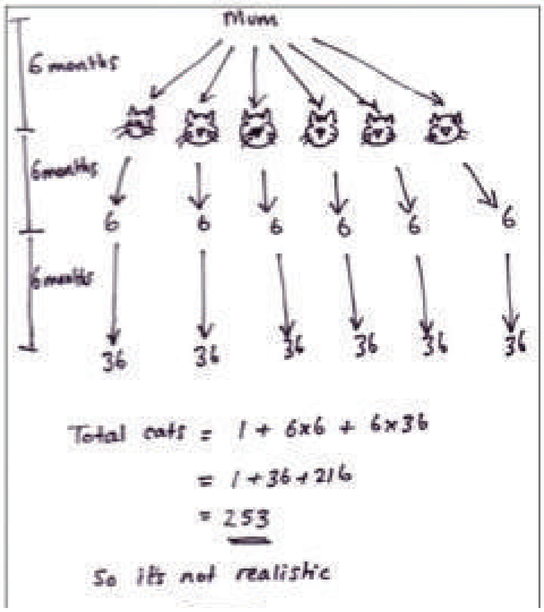
Sample Student Work



Sample Responses to Discuss: Alice


(3 litters = 18 kittens)
including mummy is 19
2nd litter will be 8 months
1st
will be able to have babys in April/born in June
2nd
will be able to have babys in August/born in October
3rd
will be able to have babys in March/May
3rd litter
Cat tree
mum
3
6
2nd litter
1st litter
Conclusion.
The mother has 18 kittens in a year each litter are 6 kittens in each.
In a year and a half the most the family will have
9846
First babys in a year yay!

Sample Responses to Discuss: Ben



Sample Responses to Discuss: Wayne

Figure 20.

a collaboration of assessment folk who already have a track record designing rich tasks and math education folk who understand what is needed.

Whether that will happen we don't know. Pushback is inevitable. There are those who fear the changes that are needed educationally on grounds of cost, complexity, litigation, and straight conservatism. We have to be ready for this and be prepared to face it—indeed the battle is on. The initial sample of candidate performance tasks from both consortia was a "don't know whether to laugh or cry" travesty. But on the hopeful side, SBAC have appointed a distinguished "math board" (including Alan) to review proposed tasks, and they have made their views clear.

Let me just finish with a little commercial for ISDDE, the International Society for Design and Development in Education. This is a body of about a hundred people, mostly professional designers of educational materials in math or science. We met last week in Berkeley. The Society's goals are: to improve standards in design, to build a design community, and to increase the impact of good design on policy and practice. We think we are having some success. The last goal is, of course, the most difficult, but you might like to look at the Society's e-journal, *Educational Designer* and, since this talk is about assessment, at the working group report on assessment design (ISDDE, 2012). This report was quite influential with both SBAC and PARCC in their strategic planning, for example in getting acceptance of WYTIWYG as an empirical fact of life, and the responsibilities that flow from this. One new concept is "strategic design" (Burkhardt, 2009). We're all concerned about the design of lessons and tasks and how they work with kids and teachers; but how the whole thing fits into the education system it aims to serve is often neglected. Such poor strategic design is the main source of low impact and unintended consequences. This is why there's so much wonderful stuff out there that nobody uses. So for any important test, ask:

Does this test assess the practices beyond novice level and encourage good teaching?

If not, fight.

References

Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education 5*, 7–74; see also *Inside the black box: raising standards through classroom assessment*. London: King's College London School of Education (2001).

Burkhardt, H. (2009). On strategic design. *Educational Designer*, *1*(3). Retrieved from: http://www.educationaldesigner.org/ed/volume1/issue3/article9

Common Core State Standards Initiative (2010). *Common Core State Standards for Mathematics*. Retrieved from http://www.edweek.org/media/25common_3.pdf

ISDDE (2012). Black, P., Burkhardt, H., Daro, P., Jones, I., Lappan, G., Pead, D., Stephens, M. (2012) High-stakes examinations to support policy. *Educational Designer*, *2*(5). Retrieved from: http://www.educationaldesigner.org/ed/volume2/issue5/article16

MAP (2012) – The Mathematics Assessment Project materials can be downloaded from http://map.mathshell.org.uk/materials/index.php where there are more detailed discussions of assessment design.

Shell Center (1987–89). Swan, M., Binns, B., Gillespie, J., & Burkhardt, H, with the Shell Center team. *Numeracy Through Problem Solving: five modules for teaching and assessment: Design a Board Game, Produce a Quiz Show, Plan a Trip, Be a Paper Engiineer*. Be a Shrewd Chooser, Harlow, UK: Longman.