That’s a Good Question
by Gerald Fitton

Multiplication,
That’s the Name of the Game,
And Each Generation,
They Play the Same.

A 1961 song by Bobby Darren.

The mathematical operation of multiplication might not have changed since it was discovered back in the mists of long ago but the techniques which are used to execute a ‘lots of’ task have ‘multiplied’ during the last couple of millennia.

The use of logarithms for multiplication was a huge leap forward. Another technique, the one I shall use this month, is called matrix multiplication. Most spreadsheets, including PipeDream and Fireworkz, have a matrix multiplication function which not only simplifies the construction of a spreadsheet but also speeds up its execution.

My worked example concerns the analysis of multi-choice questions. I used this analysis to decide whether or not a question is a ‘good’ (a useful test) or a ‘bad’ question. I shall leave some of the details until next month. Consequently, rather than getting straight to the point (do I ever?), this month I shall concentrate on the background to my example.

Screening Tests

To the amazement of many, Lecturers do things other than lecturing.

Whilst I was teaching at College one of the many ‘extra-curricular’ activities in which I became heavily involved, was screening new students for what is called ‘Numeracy’ and then later, assessing their ‘Numerical’ progress as we applied different teaching techniques.

I shall tell you something about the tests which I inherited and then developed between 1982 and 1998. I could not have developed these tests without PipeDream and my Acorn computers. At first I used an Archimedes A320, then an A440 and much later a RiscPC.

Multi-choice or Open questions

From the start I decided to use multi-choice rather than open questions.

An open question is one where the student can write down any answer they like. For example the question could be: “What is the result of adding 3 and 5?” (with no answers offered). The student can write down any answer. They are not offered a choice.

I guess that you all know what multi-choice questions look like, but just in case you don’t this is how it works. The question is the same, “What is the result of adding 3 and 5?” but the student is given a limited choice of answers such as 2, 6, 8, 10, 15, and has to tick only one box, the box adjacent to the answer they believe to be right.

I decided on multi-choice questions for three reasons.

The first is that the marking is totally objective. What I mean by “objective” is that anybody (almost regardless of their skill level) can mark the paper and all (accurate) markers will get the same number of correct answers.

The second is that answers are either ‘right’ or ‘wrong’. There is no need to look at the ‘workings’ nor to figure out the extent of the student’s knowledge or ability from their partial answer - or from their incorrect answer.

The third is a consequence of the second. A lot of papers can be marked quickly.

I pre-printed pro-forma answer sheets to give to the students. All that was necessary to mark the papers was a transparency with the correct answers highlighted. The marker placed the transparency showing the correct answers over the student’s paper. Anyone can mark a multi-choice answer sheet!

Now, I know there are a lot of you in the education business who will be against multi-choice questions for very sound reasons. Let me say that I think I know all of them but, if you insist on ‘having a go at me’ then please read to the end of my article first.

I decided on a test with five possible answers to each question.

Ranking Candidates

The first thing I did with the marked answers was to rank the students.

I can almost hear the howls of protest from those who have been brain washed into believing that such a way of labelling students is both cruel and counter productive. Of course I disagree with you but, before you look up my email address and send me a knee jerk response, please read my next month’s submission. For the want of better words, and in an attempt to avoid euphemisms I shall label those who got few right answers as ‘weak’ and those who got many right answers as ‘strong’. We offered some students, the ‘weaker’ students, additional tuition - but what did we do about the ‘stronger’ students?

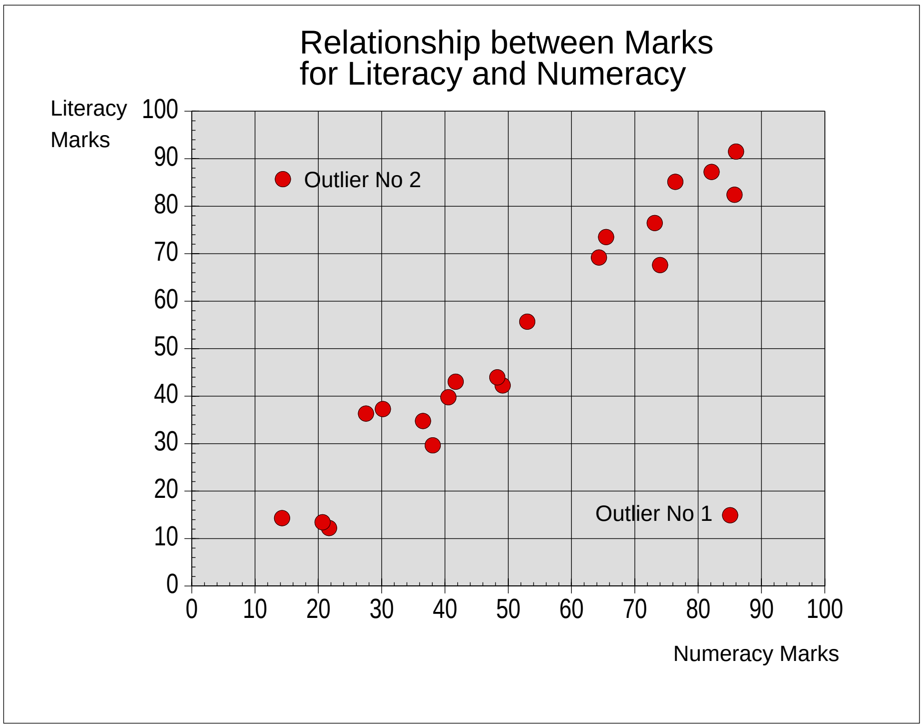
I entered all this data into PipeDream on my computer. I did not enter just the aggregate score but I included which answer (out of the five) was chosen by every student for every question. I shall come back to this shortly and then I shall tell you what I did with the results. Before I do so I have to introduce a ‘digression’!

Numeracy and Literacy

I had such success (what is success in this field? - I’ll that answer another day!) with my ‘Numeracy’ screening tests in the first couple of years (I think it was 1982 to 1984) that the course team decided that they would run a ‘Literacy’ screening test as well. Persuaded by my arguments, these were also multi-choice. They included comprehension of a written passage as well as the interpretation of advertisements and subtle differences in the meanings of words. An example of the latter is to replace the word “nice” by one which is more appropriate (from a multi-choice selection such as beautiful, colourful, green, happy).

Back to using the computer. Have a look at the drawfile below. For my own use I created very simple, basic charts in PipeDream with little additional text'; for circulation to my fellow Lecturers as part of a report, I added extra text.

The results depicted here are not for any particular year. Although they are a fictitious set they are representative. In order to simplify my presentation the drawfile shows only 22 points instead of approximately 150 points for the 150 new students I tested each year.



At first, in the early to mid 1980s, the marks in both Numeracy and Literacy were grouped in the middle range. In the 1990s the results were much more polarised; students either did very well or very badly in both Numeracy and Literacy with few students getting close to average marks. Why? Well that's another story for another day.

Correlation

The scattergraph of these fictitious results shows the marks fairly evenly distributed between 10% and 90%; this corresponds to the situation in the mid to late 1980s. You don't need to 'do the sums' to perceive that the correlation between being 'strong' at Numeracy and 'strong' at Literacy is 'statistically significant'.

Sometime around 1988 we got a new Literacy Lecturer. She decided that the Literacy test we had been using for about six years was invalid and so she set her own. It required the new influx of students to write an essay, "Why I came to Swindon College". These essays were marked by that Lecturer; the correlation with the Numeracy test was close to zero! I maintained that her marking was highly subjective (but that was not my only criticism).

We ran the old Literacy test a week later and the correlation came back to its old level. The correlation between the old Literacy test and the new test was near zero.

Of course I know the question you are asking. It is: “What are our tests are testing?” I shall sidestep from answering that just now.

Outliers

On the chart I have shown a couple of outliers which I’ve called Outlier No 1 and Outlier No 2. We always got some outliers of each type every year.

If you were calculating, say, the correlation coefficient (between the Numeracy and Literacy tests) then what would you do with these outliers? Many ‘purists’ insist that they should be included in any statistical analysis such as finding the standard deviation or correlation coefficient. Although there are cases where I would include outliers, generally I would omit them. My ‘argument’ for doing the sums without them is that these outliers are ‘samples from a different population’ and not part of the population I am investigating.

You may not be surprised to hear that those students represented by Outlier Type 1 points often had English as their second language; at home they spoke some other language.

Those of Type 2 are even more interesting.

Through additional diagnostic testing it transpired that these students had difficulty with many abstract concepts but the extent of this difficulty did vary. In the worst cases they had problems with the abstraction of colour (even though they were not colour blind) but generally the level of difficulty is that of describing the properties of (and general relationships between) abstractions such as kindness (is a certain hypothetical action kind or unkind?) and justice (what is the ‘fair’ thing to do in a hypothetical case?); relationships between abstract concepts such as Justice and Mercy could not be tackled in the abstract by these students but only in individual concrete cases (where they often got ‘right’ answers).

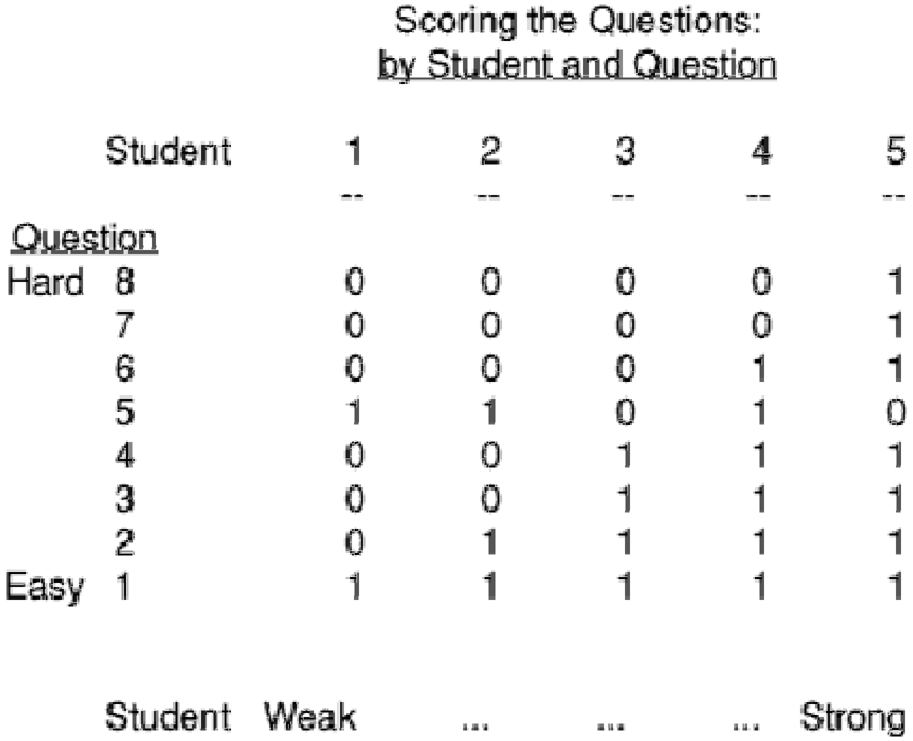
Almost invariably these students score well when they do an alternative numeracy test which includes no abstractions. 1 + 2 is an abstract sum. One apple plus two apples is concrete. These students are able to do the concrete sums but not the abstract ones!

Ranking the Questions

This is where my Archimedes (and PipeDream) really came into its own. Originally I applied the method which I describe below to study my A Level Maths students (in their second year) by using old multi-choice exam papers. It was so successful with that class that I extended the analysis to screening tests for new students of other courses.

On my Archimedes (using PipeDream) I stored the students’ answers to every question. It is relatively easy to extract from that data a table similar to that below. A “1” indicates that the student gave the correct answer; a “0” indicates that the student gave the wrong answer. These “1” and “0” are numbers so (as you’ll discover later) I can do arithmetic with them.

The table below is for illustration only. The number of students doing the test was not 5 (as in the screenshot below) but about 150. Likewise, the number of questions was not 8 but about 70. When the numbers are large (150 and 70), matrix multiplication makes the spreadsheet so much easier to set up. It would be a mammoth task without using matrices.



Scoring the Questions:
by Student and Question

Student		1	2	3	4	5
		--	--	--	--	--
Question						
Hard	8	0	0	0	0	1
	7	0	0	0	0	1
	6	0	0	0	1	1
	5	1	1	0	1	0
	4	0	0	1	1	1
	3	0	0	1	1	1
	2	0	1	1	1	1
Easy	1	1	1	1	1	1
Student		Weak	...	...	...	Strong

Let me explain to you what this table means. From the left to right of the table you will find the students in rank order with the weakest on the left and the strongest on the right. Probably you'll find that relatively easy to understand.

What you might find more difficult is the concept of ranking the questions. How is it done? What is it about the question which is measured and then ranked? The answer is that I used the students to rank the questions with the easy questions near the bottom of the table and the hard questions near to the top.

Question 1 has been answered correctly by all the students so it is an easy question. Question 8 has been answered correctly by only one student so it is the hardest question. In between I ranked the questions according to how many students got the right answer.

Bad Questions

Why rank the questions? Because it allows us to identify bad questions. In the jargon appropriate to this branch of Statistics, bad questions are called 'non discriminatory questions'. They do not allow us to discriminate between weak and strong students.

There are two bad questions amongst these eight. Can you spot them?

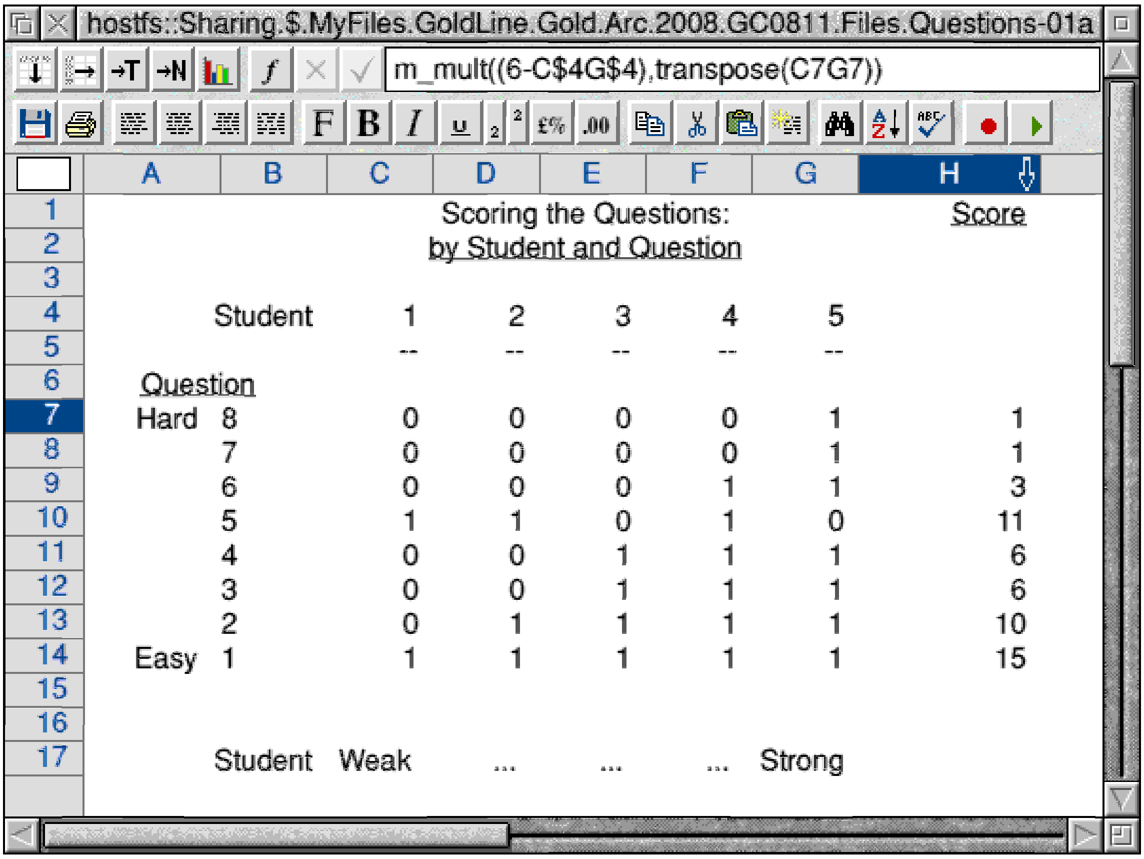
The obvious bad question is Question 1. It was answered correctly by everybody and so wasn't worth including. It does not give us any information about the relative strength or weakness of our students. Even without the painstaking analysis which I carried out using PipeDream, most testers wouldn't bother with Question 1 in future. A similar type of bad question would be one that only a statistically insignificant number can answer.

With five answers per question, a student ticking boxes at random will get 20% of the answers right so those questions for which only 20% of students answered correctly are also classed as bad questions.

Not so easy to detect as a bad question is Question 5. It is bad because it doesn’t fit the pattern set by all the other questions. When testing for Numeracy a bad question of this type might be one which asks how many people serve behind the bar in the Eastenders TV Soap Opera, how many friends they have at College or even how big their TV set is! Such a question and its answer might be interesting but it does not test the Numeracy skills of the student. Of course my suggestions as to why Question 5 is a bad question are a bit ridiculous; any reasonable examiner would recognise that such a question does not test Numeracy skills. Some bad questions, questions which do not test the skills which we wanted to test (what are we trying to test?), are harder to identify without an analysis of the sort which I have made the example of this article.

Identifying those students who can answer Question 5 correctly might be interesting but rather irrelevant. The point is that the analysis shows that Question 5 does not test the same thing as questions such as (34 − 18); subtracting 18 from 34 needs a ‘borrow’ operation. When we label Question 5 a bad question we are rejecting it because it doesn’t appear to test the same thing that the other seven questions are testing - whatever that is!

In the table I have made it quite obvious that Question 5 is a bad question. The two weakest students have answered it correctly and the strongest student has the wrong answer. Usually bad questions can not be spotted that easily. Anyway, there are degrees of ‘badness’ - the degree to which a question fails to discriminate is not qualitative but quantifiably quantitative. I may discuss the mathematics (and the construction) of a spreadsheet to quantify the degree of ‘badness’ sometime - but not in this series of articles.



hostfs::Sharing.$.MyFiles.GoldLine.Gold.Arc.2008.GC0811.Files.Questions-01a

m_mult((6-C$4G$4),transpose(C7G7))

	A	B	C	D	E	F	G	H
1				Scoring the Questions:				Score
2				by Student and Question				
3								
4		Student	1	2	3	4	5	
5			--	--	--	--	--	
6	Question							
7	Hard	8	0	0	0	0	1	1
8		7	0	0	0	0	1	1
9		6	0	0	0	1	1	3
10		5	1	1	0	1	0	11
11		4	0	0	1	1	1	6
12		3	0	0	1	1	1	6
13		2	0	1	1	1	1	10
14	Easy	1	1	1	1	1	1	15
15								
16								
17		Student	Weak	...	...	...	Strong	

The table above shows one way of spotting bad questions. It is a simplified version of the actual method which I used. In addition to the (matrix multiplication) calculation of the data shown in the screenshot for the correct answers I also analysed the ‘distractors’; my intention was to make all the distractors equally distracting - but clearly ‘wrong’.

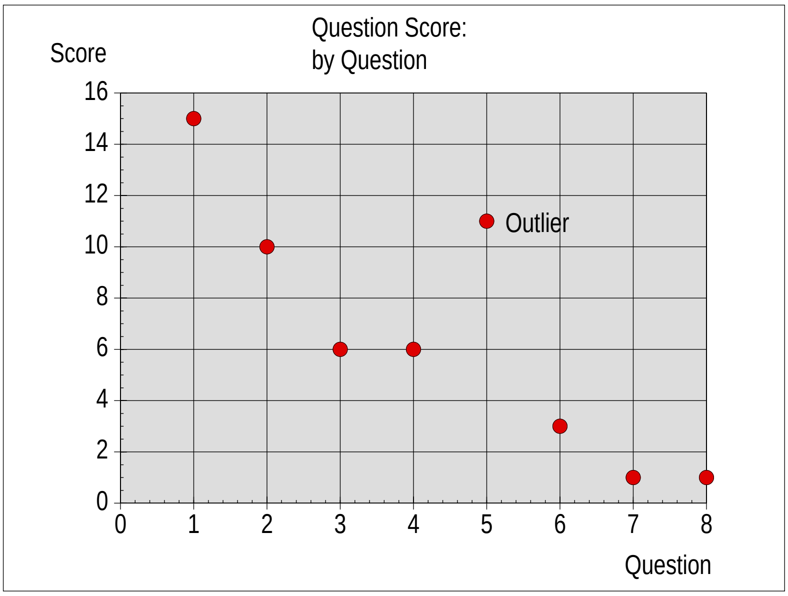
The way I arrive at a Score for each question is as follows. Those questions answered correctly by the weakest student, Student 1, are given maximum points. There are 5 Students so these questions are awarded 5 points. Those answered correctly by Student 5, the strongest Student, are given 1 point. The points scored by each question are totalled. These scores are shown in the last column of the table under the heading “Score”.

You will see that the formula in cell H7 is m_mult((6-C$4G$4),transpose(C7G7)). It is this formula, m_mult(array1,array2), the matrix multiplication of two arrays, which calculates the score of “1” returned the cell H7. The formula can be replicated down the column from H7 to H14 (to cover all eight questions) or, if there are 70 questions in the test then all the way down to the bottom of the column.

I am sure that you will appreciate that being able to replicate a formula down to the bottom of 70 rows is a lot less time consuming than typing in 70 similar formulae. What you may not appreciate immediately is that the matrix multiplication in cell H7 is equivalent to the formula (6-C$4)*C7+(6-D$4)*D7+(6-E$4)*E7+(6-F$4)*F7+(6-G$4)*G7. Now think about the work involved in writing such a formula for a row having 150 students! It is possible to write any m_mult operation ‘in full’ but the m_mult notation is much simpler.

Graphical Analysis

The graph shown below is a plot of the question score against the ranked question number.



I am sure that you will see that the general pattern of the graph is that it is close to a line from top left to bottom right.

Question 5 stands out from the general trend of the graph and I have labelled it as outlier. I used a graphical method similar to, but not identical with, this one to identify bad distractors. One such a distractor was the inclusion of 30mm when the key was 1¼ inches).

It would have been ‘nice’ (choose your own word from elegant, satisfying, perfect, complete, consistent) if Question 3 had a score of 8 rather than 6; I have some doubts about Question 8 but nothing like the doubts I have about Question 5. We have to accept that there is bound to be some ‘scatter’ about the line which represents the main trend. In fact I used a PipeDream spreadsheet to find how much deviation from the trend line could be regarded as ‘random’; beyond this range I labelled the question is a ‘bad question’.

Next time

I’ll tell you what we did with the results in my next article.

Also I’ll tell you about array operations such as adding and subtracting arrays using an array formula. Also I shall tell you more about matrix multiplication, transpose, and something which might be described as matrix division!

Box Out

An example of a bad question (included in a set of questions I had inherited from my predecessor) was one asking students to estimate the length of a nail (not a real nail but a picture of a nail) using a pair of Imperial and Metric scales (both of which were printed on the question paper). The choices were 1 inch, 1¼ inches, 25 mm, 30 mm and 35mm. The correct answer (called the ‘key’) was 1¼ inches (which is a little over 30mm) and the others (called ‘distractors’) were classed as wrong; a large proportion of students, more than 50%, chose the 30 mm distractor. I was able to discover this because I stored all the students’ answers on the Archimedes and not just whether they got the answer ‘right’ or ‘wrong’.