Draft – comments welcome

Assessing Problem Solving: Characteristics Of Student Performance On Paper-Based And Computer-Based Tasks

Malcolm Swan and Alan Bell
University of Nottingham UK with the MARS team

Introduction

Given the wide-ranging set of tasks described in the previous paper, Domain Frameworks in Mathematics and Problem solving, it is of interest to identify what general student competencies are evoked by these tasks. This paper describes some characteristics of student performance on the Problem-solving element of the World Class Tests of Mathematics and *'Problem Solving in Mathematics, Science and Technology.* These tests, recently introduced by the Department for Education and Skills in the United Kingdom, aim to provide opportunities and encouragement for achievement for pupils of high ability aged 9 and 13, and to provide an international standard of comparison. Suitable pupils will be entered individually, by parents or schools, somewhat as for graded tests in music, and successful students will receive a certificate. Though the tests are intended to present challenges to the top 10% of pupils in each age group, they are being designed with 'ramped' tasks, so that all pupils who are legitimately entered, say the top 15-20%, should have a satisfying experience.

The MARS/Shell Centre group is designing the tests of Problem solving. The materials are in two parts: computer-based and paper-based. The first administrations of the test have taken place, in November 2001 and February 2002, including entrants from Australia, New Zealand, Hong Kong, Slovenia and the USA. Previously all tasks were pretested on samples as described above. The data presented here is based on results from the pretests and the first live test. The tasks were distributed over several tests; the number of pupils taking each test varied between 100 and 200.

Initially, performance on each task was studied, looking for variations in success on the different aspects of the task. These observations were then studied to identify characteristics which appeared to have some degree of generality across tasks. This led to the points listed below. The work is thus a pilot study, rather than a definitive one. It is clear that the observations depend on the particular task set used, but the variety of task types, and the method of comparing performance on elements within each task, lead us to believe that the results are sufficiently general to be of interest. We shall continue to monitor performance in a similar way on subsequent task sets.

In quoting the evidence in the following text, the 9 and 13 year old populations are dealt with separately; but it will be seen that many of the characteristics of performance noted apply to both populations.

General

The main characteristics of performance by this group of able students are the following

1 Tasks requiring deductive reasoning and the co-ordination of constraints are better done than one might expect; by contrast, the explicit statement of justifications or explanations is at a much lower level.

2 Direct reading of common types of table and graph is well done, but success is much lower when co-ordination of aspects is required (as in recognising the relevance of the gradients of lines), or when the type of table is unfamiliar.

3 The unprompted use of any form of systematic, tabular or diagrammatic display, even when essential to the solution, is rare.

4 Testing proposed scientific hypotheses against given experimental evidence is generally good; but the making of hypotheses or mathematical generalisations from data is much harder

5 Making and expressing generalisations in mathematical pattern situations is very difficult for 9 year-olds, much easier for 13 year-olds.

6 The choice of appropriate mathematical or scientific model for a practical situation, which is itself common, but not commonly probed in this way, presents difficulty.

7 Completeness and rigour are minority attainments even in this population.

The following two points are the result of informal observation only

8 There is no clear difference in difficulty between tasks requiring spatial and numerical reasoning

9 The correct choice and performance of standard number operations is not at the high level one might expect from students at this level of ability.

The first seven points above will be supported from test data. The tasks used are described briefly in the text below. Most of them are also fully shown, in alphabetical order, in Appendix 1p (paper tasks) and 1c (computer tasks).~~ ®~~ indicates that some student responses are shown with the task.

Supporting Evidence

1 Reasoning vs explanation

Age 13

The task *Triplets*~~®~~ involves complex logical deduction from data. It presents three boys, A, B and C, in a row, each making a statement. A says: 'The one in the middle is Tom', B says: 'Hi, I'm Dick', and C says: 'The one in the middle is Harry'. Students are required to state which boy is which, and justify their answer~~®~~. The results were that **90%** correctly identified the boys**, 33%** were able to show that their own answer was a correct one, and **23%** could show also that it is the only possible answer. We see here a great difference between the success in reasoning mentally, and the ability to write down an explicit statement of the reasoning. As might be expected, there are a number who use a 'guess and check' strategy. That is, they assign names randomly then check to see if the constraints are satisfied. Such an approach does not show that the answer is the only one possible, however.

Cube Calendar (not shown) asks first for the numbering of two cubes to form a calendar showing all dates from 01 to 31, and secondly a proof that a numbering to show 01 to 52 is impossible. (Students are told that the '9' label may be turned round to make a '6').Thus, like *Triplets*, it also demands logical deduction, but in a spatial-numerical setting
The results show some similarity with *Triplets*, with a higher degree of success (**40%**) in the first part, where explicit justification is not required, and a much lower level **(11%)** in stating the argument for impossibility in the second part.

Age 9

Apples, Bananas and Pears~~®~~ shows data that seven pears weigh the same as four bananas and five bananas weigh the same as six apples, and asks for which single pieces of fruit weigh the most and the least. **35%** of students correctly identified these, but only **5%** gave adequate justifications. Throughout many similar tasks we have noted that students are unable or unwilling to give written explanations at age 9.

2 Interpreting symbols, tables and graphs

Age 13.

The use of standard Cartesian graphs and their tables of values was in general easy, except when there was a need to recognise the relevance of the gradient of a line. For example, *Rope* shows a point graph of weight against length for seven pieces of rope, and asks which points represent ropes of the same length, the same weight, and, finally, thick and thin rope; these last require the co-ordination of length & weight. **58%** of students answered the first questions correctly, but only **6%** achieved success on the final questions.

Other types of representation were much harder. For example, *Hike* gave a table in the usual triangular form for distances between five villages. This was to be turned into a diagram, and the shortest route found, from one of the villages, visiting all the rest and returning to the start. Thus the demands were first the translation of table into diagram, and then the exhaustive checking of all possible routes to establish the shortest. Only **12%** of entrants drew a correct map with distances marked; another **22%** drew maps without distances but otherwise correct. For the shortest route, **30%** identified just one feasible route, **10%** identified more than one, and only **4%** identified the correct shortest route, with some degree of justification. Thus the demand to translate an unfamiliar, but in practice common, type of distance table into a map diagram proved very difficult; and the low level of success in establishing a shortest route is another example of the great difficulty of displaying explicitly any kind of rigorous argument.

Age 9

In *Robots* (not shown) a natural symbolism for moving on a grid, such as F5, R, for forward 5 steps, turn right, presented no problem. But in a task, *Strange Rock*, which showed a 3 times table in an "ancient" notation using geometric symbols for units and groups of 4 and 4^2, **23%** of students correctly decoded two 'ancient' numbers, and **11%** successfully put two ordinary numbers into the ancient notation. The degree of unfamiliarity with this notation for very familiar objects appears to have a substantial effect. *Eggs* (not shown), concerning the distribution of coloured eggs to boys and girls, was easily solvable using a mapping diagram, but no student used any diagram. Recourse to such a method was clearly not part of their repertoire.

3 Systematic tabulation

Age 13

A number of tasks, in particular, optimisation tasks, require some systematic recording of results. *Making Soft Toys* (not shown) gives data of cost and time needed for making two types of toy, and asks how many of each should be made for maximum profit. This needs organisation of the data, and the calculation of a set of adjacent values. In the event, although many students made a number of relevant calculations, only **21%** showed any reasonably systematic approach, and only **3%** came near to a full solution. Most pupils failed to consider sufficient cases. They seemed unaware of the scale of the task or of the need for a table, graph or other systematic approach. Many went straight for the superficial response that since bears make more money, they should only make bears, thus ignoring the fact that this breaks the time constraint.

Age 9

In *Lifts,* data were given of wait time and travel time for two parallel lifts, and the final question posed was where would they pass each other. **16%** of students obtained some marks for correctly identifying floors, but only **4%** used a table with any degree of success.

In both of the above tasks, it should be noted that students are not told to construct tables or graphs. They are left to choose any appropriate approach. The unprompted use of these approaches, even when essential to the solution, is rare.

4 Scientific hypotheses and generalisations

Age 13

Pollen~~ ®~~ presented a table of 8 days' readings of data on temperature, humidity and pollen count, and asked whether the pollen count was affected by either or both of the other variables. Very few students attempted to reorganise the data using an ordered list; no graphs were drawn. These results again confirm the note above concerning the unprompted use of representations. In all, **14%** drew generally correct conclusions with some reasons, and 2% stated correct relationships with full justification.

Age 9

Skeeters tested scientific inference from data. Details were given, and pictured, of where skeeters were found in a garden – many under damp leaves, some among dry stones, very few on the path or grass. Four hypotheses were put forward, such as 'They don't like damp, light conditions'; students had to identify which hypothesis was 'not very good', and explain why. Next, an indoor experiment, recreating the four conditions of dampness and darkness, was described, with its results, and again the students were asked which of the four hypotheses were supported by these data. Finally, they were asked to explain the discrepancy between the outdoor and indoor results. **40%** of students scored 3 of the 5 marks for identifying which hypotheses were supported by the evidence, but only another **7%** scored 4 or 5. Only **20%** were able to generate an acceptable hypothesis for the discrepancy between the two sets of observations.

Balance (not shown) required placing first one, then two, given weights on a pictured balance. There were four such problems, ramped in difficulty, and finally a request to state what rules or patterns were observed. There was a great contrast in success between solving the problems, mainly done by rapid trial and error, and expressing any perceived rule. Success on the four problems was at

99, 91, 97 and 78%. On the statement of rule, **79%** scored 0, **20%** 1 or 2, and **1%** scored 3 (the third mark was given for a quantified statement, indicating, for example, that a weight twice as far away from the pivot has twice the turning effect).

5 Mathematical generalisations

Ages 9 and 13

In working with generalisations in number patterns there were large differences between 9 and 13-year-old students. At both levels there was a *Number Pyramid* task, in which a bottom row of consecutive numbers were added, in pairs, to produce the numbers in the next row above. The 9-year-old task was on paper, and had three numbers in the bottom row; the 13-year-old version was on computer, and had four numbers in the bottom row. In both cases the task was to relate the bottom left number with the top number of the pyramid. The first questions required finding the finishing number from the starter, and vice versa, in small number cases which could be worked step by step. The later questions involved recognising and stating, either verbally or by a formula, the general relationships. (They were, in formal notation, $y = 4(x + 1)$ and $y = 8x + 12$ respectively). Of the 9 year-olds, **53%** succeeded with the initial small number cases in the forward direction, but only **14%** could reverse the relation, even in numerical cases, and a mere **3%** could state the forward generalisation even verbally. Among the 13 year-olds, **85%** correctly obtained the final number from a given starter number *and* the starter for a given final number in the numerical cases, **68%** obtained a correct verbal or symbolic formula for the forward relation, and **38%** obtained a correct verbal or symbolic formula for the reverse relation. (The computer may have assisted the older students on the initial numerical questions, but not with the formulas.)

6 Modelling real situations

Age 13

Newspaper~~®~~ showed a double sheet taken from a newspaper, with its page numbers (14 and 35), and asked how many numbered pages there would be in the whole paper**.** **47%** of students failed to score; **14%** obtained a correct answer, with justification, and a further **15%** found the formula for the general case, $x + y - 1$.

Run or Swim asked students to suggest two reasons why a swimmer uses four times as much energy as a runner – a scientific modelling task. Only **16%** gave two valid reasons; **38%** gave one.

Age 9

Voting Results~~®~~ gave some facts about a class vote for the preferred book (out of three). (34 votes in all, winning book got less than half the votes, a tie for second place). Students had to find all the possible ways in which the votes could have been cast. **63%** of pupils omitted this task; **26%** stated something beyond the one result 16,9,9; of these, **7%** gave the three correct possibilities.

In both these tasks, the mathematical processing is easy; the decision about what mathematics to apply constitutes the main difficulty. In the second task, there is the added conceptual obstacle of having to think of additional possibilities, after finding the first solution, and having to be satisfied that *all* possibilities have been considered.

7 Rigour and completeness

Ages 9 and 13

There were a number of tasks in which the justification required the exhaustive checking of possibilities. *Voting Results* above, is one such; others are *Cube Calendar, Hike*, and *Triplets*, and in a slightly different way, the optimisation tasks such as *Making Soft Toys*). In all but one of these cases **11%** or (usually) fewer achieved the fully valid justifications.

Age 9

Completeness is a quality observable, if it exists, in some Design tasks. *Snakes*~~®~~ presents a small Snakes and Ladders board, containing a number of faults. The requirement is to identify and explain the faults, and to design a similar but fault-free board. **52%** of students scored 8, 9, or 10 marks out of 14, but only **2%** gained full marks – an example of how very few students achieve completeness even in a straightforward task.

Age 13

Paper Aeroplane asks students to plan an experiment to determine how the time of flight of a paper aeroplane depends on the width of its wings. They are prompted to ensure it is a fair test, to specify the measurements to be taken and the processing of the results. Only **5%** mentioned 9 or 10 of the 10 necessary points.