Computers in Mathematics Assessment

Daniel Pead
Mathematics Assessment Resource Service (MARS)
School Of Education, University of Nottingham

1: Introduction - opportunities and threats

1.1: Overview

Computers have the potential to present a wide range of mathematics assessment tasks to students, but the complications of responding to tasks via a keyboard and mouse present many difficulties, as does the analysis of the responses obtained. Furthermore, few projects will have free reign to reinvent the curriculum or make drastic changes in the style of tasks used.

This paper details recent research and development undertaken at the Mathematics Assessment Resource Service, University of Nottingham, and focusses on three different computer-based assessment projects:

- The *World Class Arena* was an ambitious programme to develop rich problem solving tasks that fully exploited the potential of the delivery medium . Although this had the luxury of being largely unconstrained by any existing syllabus or established assessment, it was constrained by the need to deliver tasks and infrastructure in quantity and on time for an internationally available test.
- Next comes an overview of MARS's evaluation of the digital version of an existing paper assessment. These tasks were far less ambitious, but faced the added challenge of dealing with very young children.
- The final example is of a small-scale project concerning just a few of the issues that might arise from computerising an established high-stakes assessment, where wholesale revision of the syllabus and assessment style would prove problematic.

The computer is, ultimately, a delivery medium and not tied to any pedagogical theory: these case studies show that solutions can be found to support – and hopefully enhance - very different assessment cultures. They also highlight many technical, practical and organisational issues and how these could, in some cases, unintentionally subvert the educational aspirations of a project.

1.2: Threats and opportunities

There are two somewhat contradictory motives for the use of computer-based assessment.

Efficiency: computer-based assessment has the potential to greatly streamline the logistics of large-scale assessment. Online delivery removes the need for printing and securely delivering tests; automatic marking removes the cost of training, employing and supervising teams of markers, allowing instant results. Together, these offer the possibility of "on demand" assessment, allowing students to take tests when they feel they are ready and freely re-take tests.

Richness and validity: the computer can deliver a greater diversity of tasks, including multimedia, simulations and adaptive tests. They offer fine control over the availability of "aids" such as calculators and reference material for each question. More information can be presented to students without undue dependence on their language skills. These features allow questions to be set in engaging contexts and using realistic data and answered using genuine ICT tools where appropriate.

Why are these contradictory? The first tends to assume short questions, amenable to simple, automatic marking. Testing on-demand, in particular, relies on large banks of well calibrated questions – which is most feasible if each question tests a single aspect of mathematical performance in isolation, on a pass/fail basis. This, coupled with statistical scaling techniques means that random tests, that produce "stable" results can be generated instantly. This approach is largely incompatible with the aims of the "functional mathematics" movement – which would argue that this item-based approach encourages a fragmented, incoherent view of mathematics.

If computers are to be used to deliver rich questions, the cost of design and development becomes a major issue. Such rich tasks are only educationally valid if they are carefully engineered, trialled and refined – it would be difficult to "commoditise" their development to allow large redundant banks to be built up. The clear experience with the "World Class Tests" project was that having a good "traditional" task designer create a paper draft and hand it over to a programmer simply didn't work.

The authoring systems used to implement "rich" tasks on computer would be more complex and require more skilled operators. The successful design of each individual task requires a challenging combination of educational, artistic and software design skills. Unlike paper based assessments, where a rough idea can be easily tested, a large proportion of the design and implementation must be completed before the task can be trialled – at which point it may fail.

It is likely that "rich" test development will be significantly more expensive than traditional paper tests, at least until more is learned and a skills base is built up. However, it is equally unlikely that an economy-led development will deliver educational improvements, and could be damaging when used for high-stakes assessment that has an inevitable effect on what is taught in classrooms.

The challenge for decision makers will be to find a good balance between the two approaches, and avoid "cosmetic" compromises (which can result in short fragmented tasks being "sexed up" with irrelevant animations and interaction). One hopeful point is that while logistical considerations mean that a traditional test paper has to be *either* multi-choice or constructed response; calculator or non-calculator, a computer-delivered test can be more heterogeneous, and could easily mix a battery of short no-calculator-allowed arithmetic questions with more extended tasks.

Another major issue is the state of ICT provision in schools – in the UK great progress has been made in equipping schools with computer labs and internet connectivity. However, this still falls short of the needs of large-scale computer based assessment. The problem is not just the number of "seats" but the priority given to support and maintenance – if computerised testing is to be widespread then ICT provision in schools will become "mission critical" necessitating the sort of redundant systems and fast-response support used by IT-dependent businesses.

Clearly, the "testing on demand" principle is partly motivated by the ICT provision issue, but does not take into account the practical classroom management issues of dealing with students at different stages in the assessment process.

Above all, it should be remembered that the content of high-stakes assessment, combined with pressures for school "accountability" has an enormous impact on daily classroom teaching. E.g. *Shepard, 1989* noted that teachers model their lessons on the *format* of the test, not just the topics covered. The responsibility of test designers – in any medium – extends far beyond simply delivering stable and reliable assessment.

2: Case study – World Class Arena

2.1: The product

The World Class Arena was a QCA/DfES funded programme to identify and engage potentially gifted and talented students, particularly those who's ability was not apparent from their performance on day-to-day classroom activities. The project focussed on two related subject areas - "Mathematics" and "Problem solving in mathematics, science and technology". Most of the work on the former was done by the AEU at the University of Leeds. This article will focus on the "Problem solving" materials, which were designed by the Shell Centre/MARS team at the universities of Nottingham and Durham.

A key constraint of the design – which has resonance with some models of functional mathematics/mathematical literacy – was that the tasks should *not* require above-average curriculum knowledge, but should focus on more sophisticated reasoning and insight (e.g. *Steen, 2000*).

The product, as originally conceived in 1999, would consist of computer-delivered assessment tests for students at ages 9 and 13. The original concept called for four sittings a year – but as the project continued it became clear that this was surplus to demand and unsustainable and the number of annual sittings was reduced to 2 and the production of classroom support material was introduced.

Another decision made shortly after the project started, based on the consensus of both teams of designers, was that the "state of the art" of computer-based testing – particularly when it came to response gathering – would limit the opportunity to cover the full domain of "problem solving" - which, in itself, was relatively unexplored, so only half of the test would be on computer, the other half would be pen-and-paper.

More contentious, was that the computer-based tests would also partly rely on paper answer books. The latter – though probably untenable in the long term – did provide a valuable stepping stone. Towards the end of the project, as experience was gained, the dependence on the answer books was waning and, had task development continued, the answer books would probably have been dropped, or relegated to "rough work" which would not be marked.

The availability of the written paper meant that the computer tests did not have to waste effort replicating tasks that were known to work well on paper, and could concentrate on ideas that exploited the computer to the full.

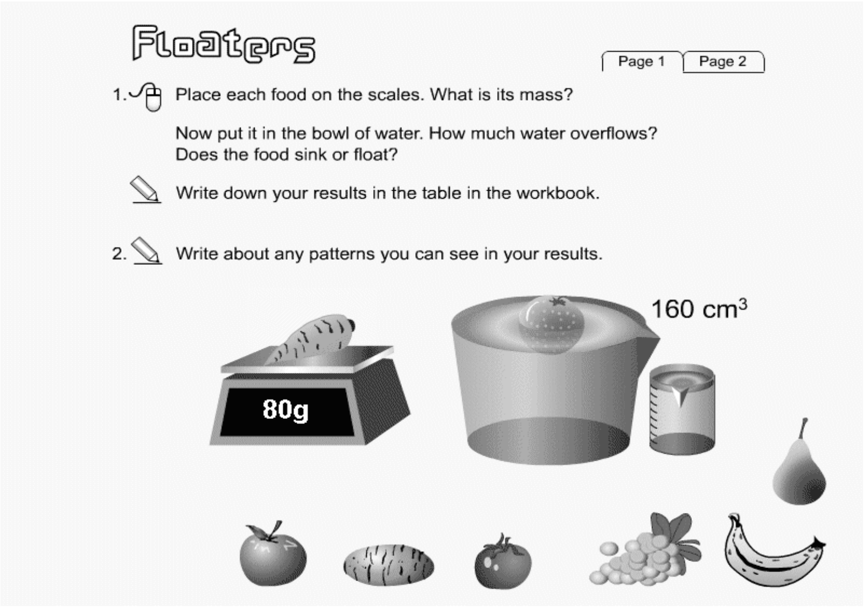
2.2: Styles of question

Simulated experiments and microworlds

One of the challenges for the problem solving strand was to cover the field of "problem solving in science" without (as discussed above) depending on advanced curriculum knowledge – a particular problem at age 9. The computer allowed the presentation of simulated science experiments – in a simplified but defensible form – that embodied all the required data and left students to investigate, draw inferences and justify their conclusions. Figure 1 shows one example, which had 9-year-olds successfully engaging with the beginnings of Archimedes' principle, eliciting insightful responses such as

> *"All the vegetables and fruits that sinks overflow less than they way. All the food that float overflow how much they way"*

(Our job here is done – the rest is up to the English department!)



Floaters

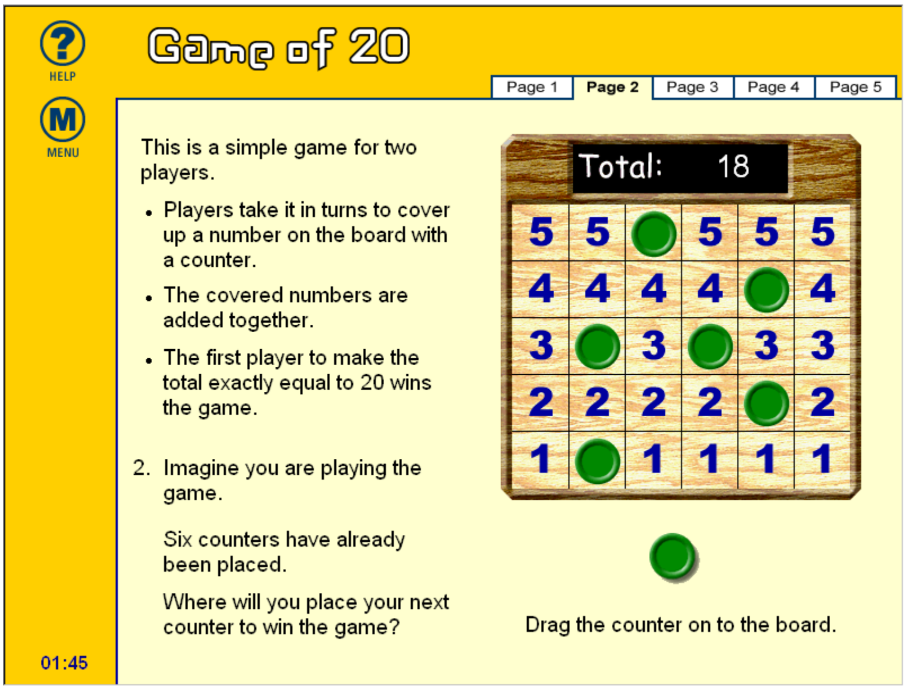
1. Place each food on the scales. What is its mass?

Now put it in the bowl of water. How much water overflows?
Does the food sink or float?

Write down your results in the table in the workbook.

2. Write about any patterns you can see in your results.

Figure 1: Floaters - a simulated science experiment



This is a simple game for two players.

- Players take it in turns to cover up a number on the board with a counter.
- The covered numbers are added together.
- The first player to make the total exactly equal to 20 wins the game.

2. Imagine you are playing the game.

   Six counters have already been placed.

   Where will you place your next counter to win the game?

Total: 18

5	5	●	5	5	5
4	4	4	4	●	4
3	●	3	●	3	3
2	2	2	2	●	2
1	●	1	1	1	1

●

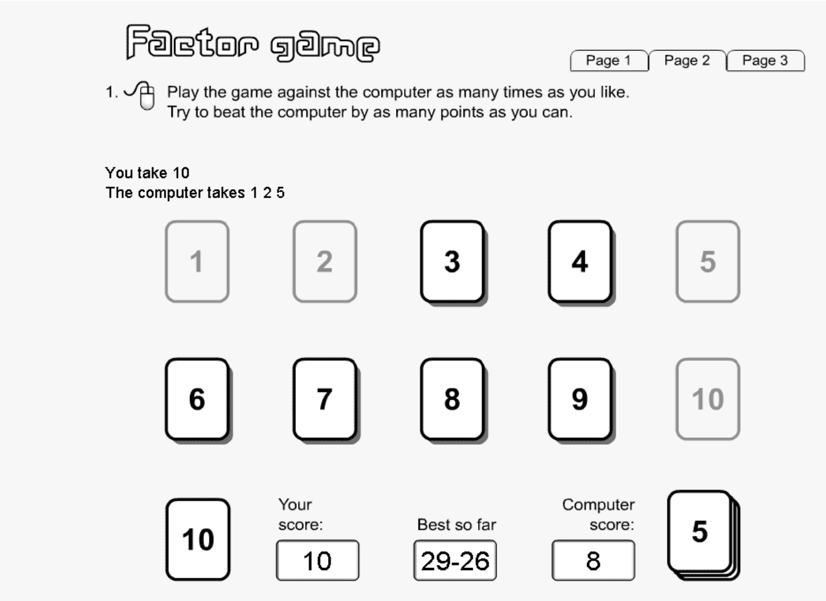
Drag the counter on to the board.

Figure 2: Extract from the "Game of 20" task

The tests were not limited to “real life” problems and included several “Number games” such as the example in Figure 2. This type of game (a variant of “Nim”) has the advantage that there is an easily accessible optimum strategy. However, it was soon clear that leaping directly to formulating the strategy was beyond the ability of most students, so these tasks typically fell into the pattern:

1. Here are the rules – play a few games against the computer.
2. Here is the last stage in a sample game – identify the winning move
3. Here is another sample game – identify the two moves needed to win
4. Now describe the strategy for always winning the game.

In “Factor game” (Figure 3) the computer played a key role in explaining the rules of the game using an animated sequence. The student's ability to formulate a strategy was put to the test by challenging them to beat the computer by the biggest margin possible. As a follow up, their understanding of the strategy was probed by asking them to select the first two moves in a (non-interactive) variant of the game with 50 cards.



1. Play the game against the computer as many times as you like.
Try to beat the computer by as many points as you can.

You take 10
The computer takes 1 2 5

Your score: 10

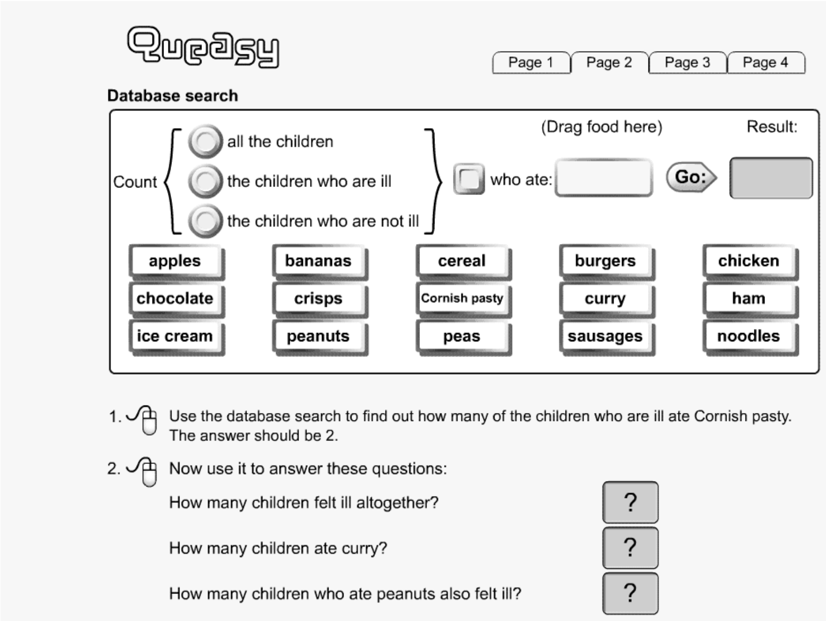
Best so far: 29-26

Computer score: 8

Figure 3: Factor game - student vs. computer

Exploring rich data sets

One advantage of computer-based tasks is that the student can be presented with a substantial database, rather than the dozen-or-so cases feasible in a paper test. *Queasy* (Figure 4) requires students to solve a food mystery by making suitable queries to a simulated database while *Water fleas* (Figure 5) allows a large set of experimental results with several variables to be viewed as bar charts and asks whether these results support or refute a series of hypotheses.



1. Use the database search to find out how many of the children who are ill ate Cornish pasty.
The answer should be 2.

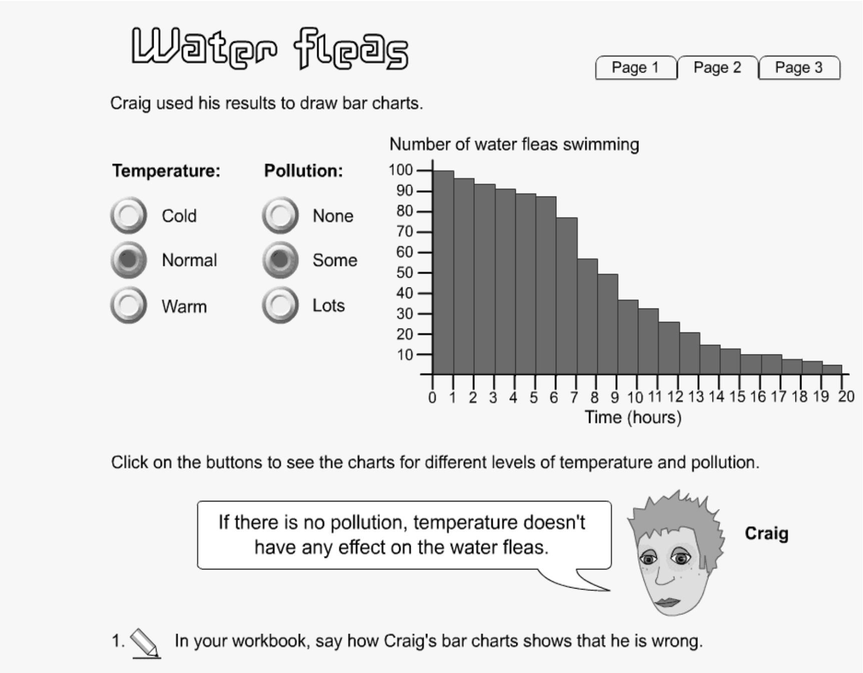
2. Now use it to answer these questions:

How many children felt ill altogether? ?

How many children ate curry? ?

How many children who ate peanuts also felt ill? ?

Figure 4: Exploring a database



Water fleas

Craig used his results to draw bar charts.

Temperature: ○ Cold ● Normal ○ Warm

Pollution: ○ None ● Some ○ Lots

Click on the buttons to see the charts for different levels of temperature and pollution.

If there is no pollution, temperature doesn't have any effect on the water fleas.

1. In your workbook, say how Craig's bar charts shows that he is wrong.

Figure 5: Water fleas

2.3: Some design challenges

Finding the task in the context

The desire for rich and interesting contexts has to be balanced with the constraints of the assessment. Various ideas emerged from brainstorming sessions – such as Muybridge's famous pictures of galloping horses, or analysis and comparison of demographic data from many countries – with no notion as to what a self-contained 5-15 minute task (with an outcome that could be methodically marked) might be.

One of the hardest decisions for a lead designer was when (and how) to diplomatically reject a contributed idea, into which a lot of research and effort had already been put and which would make a wonderful extended investigation, on the grounds that no clear task had been identified.

Eliminating trial and error

The examples above illustrate one of the key challenges of designing interactive test items – having presented the simulation, how do you probe the students' understanding of the situation in a way that can't be satisfied by trial and error. Possible approaches include:

- Written explanation – describe your strategy/justify your findings, or support/refute some suggested hypotheses.
- Simple challenge – ask them to beat the computer and rely on the "time penalty" to discourage brute force/trial and error solutions.
- Logging and analysis – record every interaction between the student and computer and then try to analyse this data to spot promising patterns and sequences. This

requires complex coding and could be fragile – a few random interactions not indicative of the students' thought processes could disguise a valid response.

- Heuristic inference – with "factor game" the final score was taken to be indicative of the level of understanding. Most students could beat the computer eventually; a high score of 30 suggested that the student grasped the idea of factors and multiple; 35 implied the beginnings of a strategy and it was assumed that the optimum score of 40 was unlikely to be achieved without a well developed strategy. This has the advantage of being easy to program and fairly easy to defend – but it might not always be applicable.
- Extension problems – also used in factor game, after the interactive session the student is asked to use their understanding of the scenario to make an inference or prediction about an extended or generalised variant, with no simulation available.

Educational design to software design

A successful task relies on both creative and pedagogically sound ideas from an experienced assessment designer, and the skilled implantation of this within the capabilities and constraints of the delivery system. Most importantly – unless both skill sets are available in the same person – it requires communication and cooperation between the two. It became painfully obvious during the work on the World Class Arena that this communication channel is difficult to establish – partly because of the widely different sets of concepts and terminology involved, but also because of organizational structures.

2.4: Some issues with the project

The early years of the project were somewhat fraught, and there may be some lessons to be learned for future projects. Some of the issues included:

- **Structure of the project** – the organisation, as conceived, was heavily compartmentalised – with two groups contracted to work on the educational design, a third contractor handling the software development and a fourth (appointed later) responsible for "delivering" the tests. This seemed to be founded in a publishing metaphor: manuscript -> editor/designer -> printer -> distributor and, initially, failed to take account of the need for the designers' aspirations to match the capability and resources of the programmers.
- **Technical oversight –** the project had several stages of internal and external review to ensure the educational validity of the materials. There was, initially, no corresponding oversight of the technical issues or agreement between the designers and programmers as to what the constraints or expectaions of the system were. An internal committee was eventually set up, but its source of authority was rather unclear.
- **Timing** – although the overall timescale – 2 years until the first live sittings - was appropriate, the contract called for a large scale trial a few months after the effective start of the project – and much time and effort was spent trying (and failing) to get a workable IT system in place for that.
- **Short-term contracts & rights** – this affected the programming side in particular – with no on-gong commitment to continue the contract after the initial 2 years and all IP rights assigned to the client, there was little commercial incentive to invest time in building a solid IT infrastructure which could be handed over to the lowest bidder at

the end of the term.

2.5: Outcome of the project

Development of new test items was stopped in 2003, but test sittings continue with the existing materials – see www.worldclassarena.org. From that site: “Since the first test session in 2001, over 18,000 students in over 25 different countries worldwide such as Australia, Hong Kong, New Zealand, Saudi Arabia, Slovenia, the United Arab Emirates, the United Kingdom and the United States have taken the tests.”

The classroom materials are available from nferNelson, including 6 modules under the title *Developing Problem Solving.*

More details of the design philosophy of these tests can be found in *Burkhardt, Pead, 2003*

3: Case study – GCSE

The World Class Arena project had one huge advantage in that there was no strongly established curriculum or traditional assessment for problem solving. As a follow-up to this, it was decided to investigate the possibilities of presenting GCSE[1] questions on computer.

Ambitious plans are afoot to computerise large-scale, high-stakes assessment - in the UK[2], a recent initiative from the QCA[3] states that, by 2009, “all existing GCSEs, AS and A2[4] examinations should be available on-screen” *Boston 2004.* Although there is continuous “incremental” evolution of the GCSE syllabus, it seems unlikely that a radical overhaul would happen in that time scale – so there would be a need to translate many existing question types to computer.

3.1: Rich questions at GCSE?

One of the motivations behind this study was to see if the introduction of eAssessment would change the nature of the current Mathematics paper-based GCSE. Compared to the mainly multiple-choice questions used in other countries (particularly the USA) the England and Wales GCSE papers seem relatively "rich" - they use written answers; provide space for working and allocate some marks for method. Questions may have multiple related parts and are often set "in context" rather than being presented as raw mathematical exercises.

However, looking more closely at a number of recent papers (with a view to computer-based presentation) reveals that much of this is superficial.

- It is usually explicit what mathematical technique is required for each question - questions are often “generic” types seen year after year with minor variations
- “Method marks” are usually awarded automatically if the final answer is correct. There is rarely any requirement to show or explain working except as an insurance

[1] GCSE is the main qualification at age 16, the end of compulsory education, in England and Wales

[2] For the purposes of this document, UK refers most reliably to England and Wales. There are some significant differences (such as a different examination system) in Scotland and Northern Ireland.

[3] The Qualifications and Curriculum Authority for the UK – the body responsible for the national curriculum and standardisation of examinations.

[4] In England and Wales, "A Level" is a 16-18 qualification used for university entrance and higher-end vocational purposes, now split into two stages, AS and A2

against error.

- Although the overall weighting of questions may be appropriate, it is questionable whether the partial marks represent a valid progression towards the task goal
- The “longer” questions are often a series of independent short questions sharing a context or a diagram. There are few "extended chains of reasoning".

Consequently, one question to be addressed is whether the additional effort required to collect working and award partial marks is – given the other constraints on this assessment – a worthwhile effort.

3.2: The study

The study involved the design and production of a prototype online eAssessment delivery and marking system, the design of a number of groups of “equivalent” tasks presented on paper and at two levels of computerisation. This system was then trialled with 14-15 year old students at a number of schools.

The questions to be addressed were:

- How can existing short- and constructed-answer assessments be translated into eAssessment without compromising the assessment goals
- How can eAssessment be designed to present richer tasks and to capture and mark richer responses?
- How does this add to the summative, diagnostic and formative value of the assessment compared with (a) final-response-only marking and (b) written answers.
- What are the long-term implications for the syllabus?

The two forms of “rich” input tried were chosen to enable major sub-genres of GCSE-style tasks – a “printing calculator” that could be used to capture method in computation-based tasks and a graph/drawing tool for scatter- and line- graph questions

3.3: Some findings

The major hurdle in this study was recruiting sufficient schools to produce a balanced sample – the study relied on a “cross-over” design in which students sat one of the two task sets on paper and took the other on computer - several schools dropped out after the papers had been allocated and upset the mix of papers. The ability range of students entered also varied – reducing the sample size of comparable students.

The initial finding was that there while was no clear evidence for an overall effect, the performance on individual tasks could be strongly influenced by the presentation. For example, insisting on working for a multi-step calculation seemed to improve results – possibly causing students to work more carefully, but if the answer was easy to intuit (such as the difference between two numbers in a table) it could have a distracting effect.

Analysis of the results is continuing.

The trials also revealed that some students at this age are highly skeptical about the value of computer-based testing at this level – a contrast from the younger students involved in the other case studies. The prize quotation was:

> *...also whats wrong with paper. and manual labour. i dont know what is trying to be proved by using computers, but mine started flashing purple and things and went fuzzy and put me off from answering questions. this WAS NOT HELPFULL you made me very stressed, although it did make me chuckle.*

...a number of candidates made less loquacious criticisms, some attributable to specific problems with the prototype system, others (as above) expressing more general skepticism and zero tolerance for bugs.

It was also clear that school infrastructure is not yet able to cope with widespread computer based testing, particularly with regard to the speed and stability of internet connections.

> *"Our school has one 1 meg* (Mb/s) *connection for 1500 people (As you know, schools have no choice of ISP)"* [5]
>
> *"One of the e-desk rooms was invaded by plumbers 20mins before we were due to start and we had to relocate. We had a power cut 5mins before we started."*

3.4: Some tools for gathering rich input

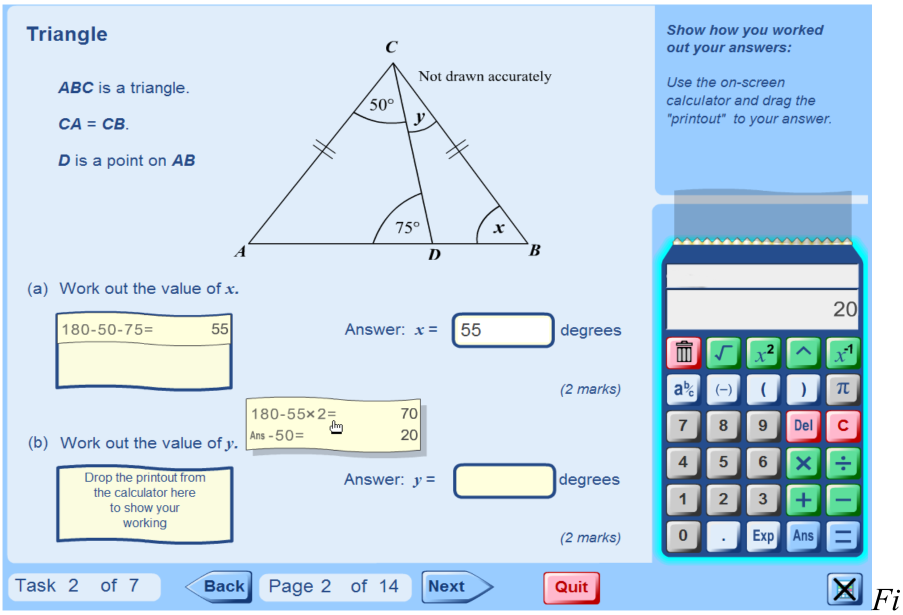
The "Printing calculator"

On a pair of past GCSE papers analysed, 22% of the marks could potentially be awarded as partial credit if a particular number or calculation was seen in the students working. One problem from the computer assessment point-of-view is that this working is "free form" - just some ruled lines and the standing instruction to "always show your working". Typing sums into a text box might not be a natural way to work.

Since a "natural medium" for doing this form of mathematics would be a pocket calculator, an on-screen calculator could be useful for capturing such working. However, building on the experience from World Class Arena that deducing students intent from interaction log files is complex and fragile, it was decided to put the onus on the student to decide what to exhibit as their working.

The result was the "printing calculator" - Figure 6 - an on screen calculator featuring a printout that can be "pinned" to the students response to show working.

[5] 1 Mb/s is typically the entry-level "home broadband" connection offered for ~£20/month – with 2-8 Mb/s being available in many areas – admittedly this is a "maximum" speed and business users might pay more for guaranteed bandwidth. Note 1 Mb/s (bits) is approx 100 MB/s (bytes).



Triangle

ABC is a triangle.

CA = *CB*.

D is a point on *AB*

(a) Work out the value of x.

180-50-75= 55

Answer: x = 55 degrees

(2 marks)

180-55×2= 70
Ans -50= 20

(b) Work out the value of y.

Drop the printout from the calculator here to show your working

Answer: y = degrees

(2 marks)

Show how you worked out your answers:

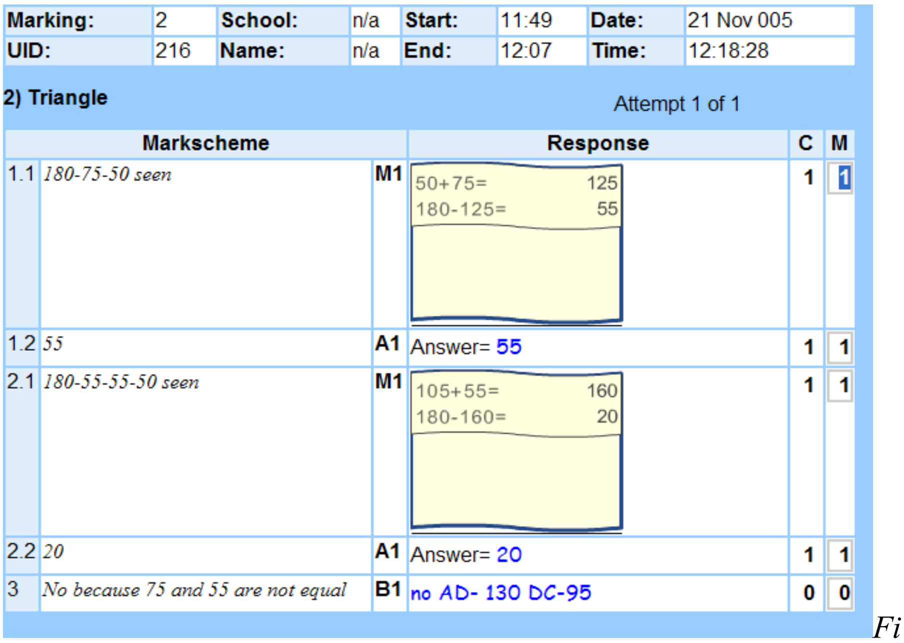
Use the on-screen calculator and drag the "printout" to your answer.

Figure 6: The "printing calculator" in use

The markers are presented with a screen similar to Figure 7 - although, for the trials, the "C" column indicating the marks awarded by the computer was hidden and the markers worked independently.

This illustrates an issue with this type of mark scheme when a calculator is used – it is relatively unlikely that a student will enter the correct calculation and then get the wrong result – yet this mark scheme is fairly common in GCSE (even on the paper where a calculator is allowed), where "method marks" are usually awarded by default if the final answer is correct. However, on occasion a question needs to refer to the working – for example, if there is a yes/no answer that needs to be distinguished from a lucky guess.

An interesting (but tentative) result from the trials of this question was that more students got both (a) and (b) correct when the printing calculator was used, compared with a version of the question in which they simply entered the answer (but still had access to a calculator). Other questions showed the opposite effect – suggesting that being forced to supply working can either help or hinder, depending on the context.



Marking:	2	**School:**	n/a	**Start:**	11:49	**Date:**	21 Nov 005
UID:	216	**Name:**	n/a	**End:**	12:07	**Time:**	12:18:28

2) Triangle

Attempt 1 of 1

	Markscheme		**Response**	**C**	**M**
1.1	*180-75-50 seen*	**M1**	50+75= 125 180-125= 55	1	1
1.2	*55*	**A1**	Answer= 55	1	1
2.1	*180-55-55-50 seen*	**M1**	105+55= 160 180-160= 20	1	1
2.2	*20*	**A1**	Answer= 20	1	1
3	*No because 75 and 55 are not equal*	**B1**	no AD- 130 DC-95	0	0

Figure 7: Marking the calculator responses

Graphing tools

Another common genre of GCSE question requires the drawing and marking of line graphs – such as plotting equations or drawing a line of best fit. Figure 8 shows an example task, requiring students to draw a line of best fit and identify two errant points. The computer was able to mark this type of question by checking the gradient and position of the line.

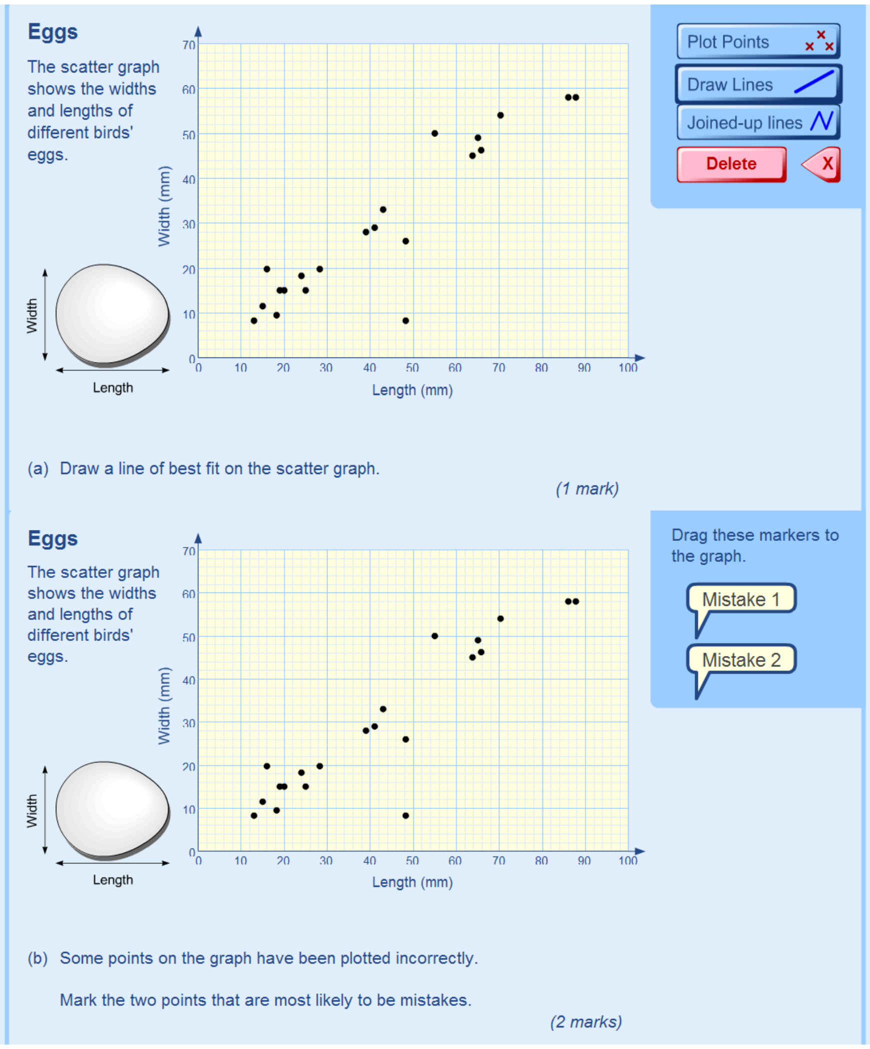
Like most computer drawing systems, the graph tool required the student to click on the beginning and end points of the line. This is a rather awkward way of drawing the "line of best fit" compared to a ruler, although students were able to use the graph tool quite successfully.

The graphing tool could also be used for any task requiring a simple, rectilinear drawing.

3.5: Marking

Some, but not all of the questions were amenable to computer-based marking. The trial results were scored by experienced GCSE markers using an on-line marking system that presented the mark scheme and students' responses side-by-side. Some examples can be seen in the illustrations in this section. The system was designed so that the computer could mark everything possible and the markers would "approve" the computer marks, and mark any sections that the computer couldn't handle. For the trial, though, the computer marks were hidden and the markers worked independently. The system worked well – after a few initial tweaks the markers quickly mastered the system and made rapid progress. This may have been helped by the use of GCSE-style conventions in the mark scheme.

Encoding the mark schemes was challenging – the GCSE conventions are quite involved, with some points being automatically awarded if others are given, follow-through of mistakes and other details. The computer language developed to encode the schemes became quite complex to allow (for example) the mark scheme for one "point" to reference the answer to a previous question.



Eggs

The scatter graph shows the widths and lengths of different birds' eggs.

(a) Draw a line of best fit on the scatter graph.

(1 mark)

Eggs

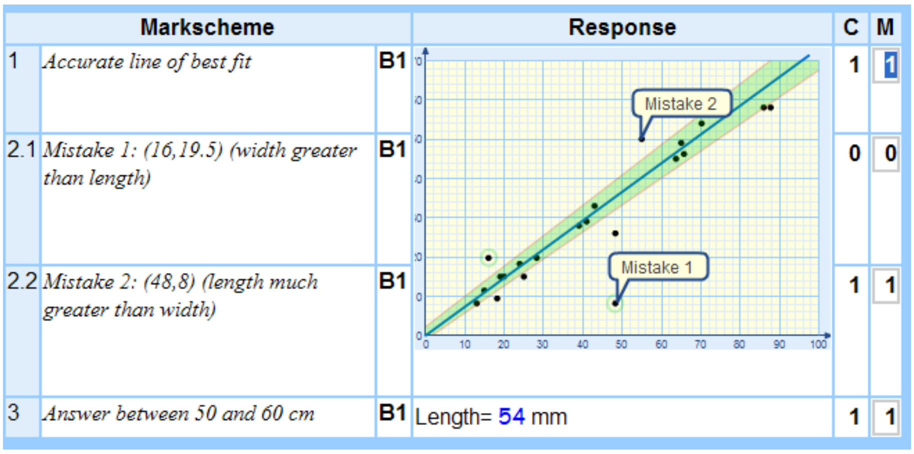
The scatter graph shows the widths and lengths of different birds' eggs.

(b) Some points on the graph have been plotted incorrectly.

Mark the two points that are most likely to be mistakes.

(2 marks)

Figure 8: Task using the graphing tool and draggable labels



	Markscheme		Response	C	M
1	*Accurate line of best fit*	B1		1	1
2.1	*Mistake 1: (16,19.5) (width greater than length)*	B1		0	0
2.2	*Mistake 2: (48,8) (length much greater than width)*	B1		1	1
3	*Answer between 50 and 60 cm*	B1	Length= 54 mm	1	1

Figure 9: Marking the task from the previous figure

4: Case study – Younger children

4.1: Background

In 2006, the MARS group undertook a review of some new computer-based tests for 6,7,8 and 11 year olds for a major UK educational publisher. The question was, whether the tests were comparable with the previous paper versions on which they were based – a trial conducted by the publisher had suggested that the on-line versions were proving slightly harder.

This study consisted of a statistical analysis of results from the publisher's own "equating study", a critical review of the task design and close observation of individual students working on the test.

4.2: The problems of conducting trials

One issue that was quickly noted was that the equating study had – as with our own GCSE trials – suffered from sample distortion due to school issues. In this case, there was no "cross-over", but students took both the paper version of the test and the (substantially similar) on-line version. However, it had not been possible to control the order in which the tests were taken or the interval between them – with some schools taking both on the same day. Although there appeared to be a strong effect from the order of sitting, this turned out to be associated with just one or two schools showing disproportionately bad on-line results. However, even eliminating these suggested a consistent underperformance on the on-line tests.

4.3: Mode of presentation

One distinguishing feature of these tests were that the paper versions for 6,7 and 8 year-olds required the teacher to read out the questions – the printed booklets had very minimal text. The on-line versions used a recorded voice.

Interviews with teachers who had administered the paper test revealed that most – while working within the rules of the test – were pro-active about ensuring the students paid attention and remained on task by, for example, repeating the question several times. In the on-line version, students worked at their own pace and only heard the question once, unless they actively chose to have it repeated. Observations of the students taking the on-line test suggested that failing to pay attention to the recorded voice was a major cause of mistakes.

This does not invalidate the tests – but does mean that the on-line versions are less forgiving of poor attention spans, and could easily account for the observed discrepancies.

4.4: Effects of question design

Most of the items in these tests were fairly faithful representations of the paper versions, which predominantly required multi-choice or short numerical answers that were easy to capture.

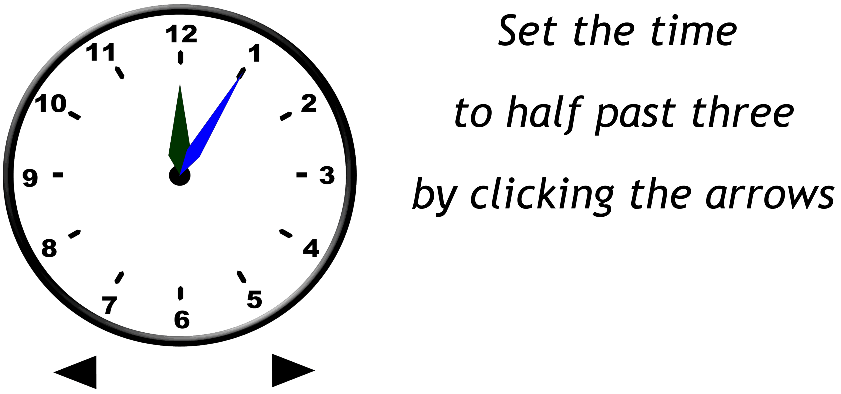
A few were more problematic, and one question type in particular highlighted the need for rigorous small-scale trials and observations of any new interface.

Figure 10 shows the style of question involved – the paper equivalent simply provided a blank clock face on which students' drew the hands. In the on-line version, the clock was moved back and forth in 15-minute steps.

In the observations, it was clear that a significant number of students *who know the correct answer* failed this question because they had problems setting the on-screen clock. When handed a “cardboard” clock or asked to draw the face they could answer correctly.

The two main problems seemed to be:

- Many students at age 6 had clearly learnt to tell the time using the “the little hand points to...; the big hand points to...;” mantra. Because the computer linked the modtion of the two hands, students would set one hand before turning their attention to the other – and then get confused (or fail to notice) when the first hand didn't stay where they put it.
- The computer represented the movement of the hour hand accurately – so, at 3:30, it was between 3 and 4. This clearly confused students who wanted to point the hour hand at 3 – an answer which would have been accepted, at this age, on paper.



Set the time

to half past three

by clicking the arrows

Figure 10.: Set the time by clicking the arrows

This case is interesting in that the computer is not “wrong” - arguably, this is showing a weakness in the way students are taught to tell the time.

Other – less dramatic – questions raised included:

- whether clicking the correct bar on a bar chart was actually equivalent to reading the label for that bar and writing it on an answer line
- whether answering a “difference” question by drawing the correct number of objects was harder than adding the objects to the screen by clicking an “add” button repeatedly – making it easy to “count on”
- whether a multiple-choice paper item showing all the possible responses side-by-side was “fairer” computer item which cycled through the possible responses in sequence – meaning that a child stopping at one of the “distractors” might never have been shown the correct response.
- whether the interactive elements introduced in some new question types actually contributed to the assessment objectives – some questions seemed to require the student to drag a pointer or click a button to reveal essential information which could easily have been shown on screen
- whether splitting a question over two screens affected the difficulty. There was some

evidence that – if the second part did not depend on the first answer – more students correctly responded to the second part when it was presented as a separate question

4.5: Distractions

The tests revealed how easy it is to divert young children's attention away from the task in hand. The tests included several attractive background screens used while instructions were being read out by the recorded voice. At the end of the instructions, some children were clearly unable to proceed (which usually just entailed clicking "next") and, when asked what they thought they should do next, talked about the picture rather than the instructions.

The number of extra instructions needed for the computer test were notable – most questions had some variation of "click in the box and type your answer" or "drag the ... to the ..." in the prompt where the paper version simply said "write..." or "draw...".

Another form of distraction arose because the children were able to start manipulating the question before the question had been read out completely. More able students often anticipated the question correctly, but others were then "tripped up" by the question. In one case, by the time the question said "add the missing dots" the student had added some dots and was unable to identify the "missing" ones.

4.6: Usability by young children

Few of the children showed any serious difficulty in operating the computer – although sometimes they were painful to watch. Some issues about the suitability of equipment are mentioned below.

There was only one question on the test that required alphabetic text entry – an initial "type in your name" screen – which some children were very slow to complete; wider use of alphabetic text entry at ages 6-7 could be a problem. Few problems were seen with typing in numbers, although analysis of the equating study results suggested a slight tendency to type numbers back-to-front (which you might do on paper when adding "tens and units").

4.7: Attitudes and reactions

The children were almost uniformly positive about taking these tests on computer – although it was also apparent that eliciting reflective opinions from 6 and 7 year olds without "leading" them is a difficult task.

The small proportion of dissenting voices (about 3 students out of the 70 interviewed) raised the following issues:

- *"I like writing"*
- *"It was easier to read on paper"* - this raises the complex issue of catering for various types of reading difficulty, and whether or not choices of background colour etc. would help.
- *"Harder to concentrate – sound was distracting – speech was too slow"* - this corresponds to points noticed during the observations. It was also apparent that the spoken prompts could be frustrating for the more able students, who would have easily coped with printed questions.

Bearing in mind that not all of the younger children were sufficiently articulate to communicate such views, these issues could be affecting a larger group of children would bear further investigation.

Teachers were also enthusiastic about the tests, the overwhelming concern being that their school would not have the capacity for whole-class sittings.

4.8: ICT provision

The latter is a particular problem with the youngest children. The schools visited were a mixture of dedicated infant schools (with just the younger children) and "primary" schools with students aged 5-11. Only one of the "infant" schools had a dedicated computer room that could seat a whole class - the rest had a small number of computers in each classroom (and, in some cases, were seen to be making excellent use of integrating these into lessons). Several were hoping to get a class set of laptops in the near future, which would be their best hope for using these tests.

In the case of the "primary" schools, computer labs *were* available, but tended to be designed with the older students in mind – in one case using high stools that were totally unsuitable for 6-7 year-olds. Again, teachers of the younger classes tended to use whiteboards and in-class PCs.

Although the in-class PCs in most schools had suitable seating, few other concessions were made to small hands – apart from a few keyboards with "handwritten" letters the equipment was full sized.   In particular, although there were no complaints, it was clear that "child-sized" mice would have been more convenient for the children. There were also some occasions where viewing LCD screens from awkward angles may have been causing colour shifts.

5: Some conclusions

5.1: Motivations

Computers are only one of many tools. Beyond an initial honeymoon period where the novelty of using computers provides motivation and engagement (a honeymoon which appears to be over for the 14-15 year-olds in the GCSE trial) mere use of ICT will not guarantee educational improvements.

It is essential – early on in a project – to identify what it is that computer-based assessment is expected to add to the pedagogical value of the assessment. Possibilities are:

- Richer, more realistic tasks
- Instant results/testing on demand
- Logistical efficiency and economy

It should be recognised that these aspirations are not entirely compatible. While economy should not be dismissed as a motivation, the wider costs of ICT provision must be considered.

Producing richer, more realistic tasks promises great improvements in the validity of assessment – but it entails innovative development with extensive trialling and refinement, and also relies the willingness to reconsider the underlying curriculum/syllabus.

Combining instant results, testing on demand with richer tasks may be possible, but this requirement is liable to create a strong economic pressure towards more simplistic assessments. The implications of "test on demand" for day-to-day classroom management should also be considered.

5.2: Impact on validity

There is a widespread suspicion that testing mathematics on a computer can lead to a small decrease in performance compared to conventional tests. Our studies on the 6-7 year old tests appeared to show a consistent trend, albeit of marginal significance, even when a few problematic items and schools with possible technical issues were excluded. The study of GCSE-style questions had inconclusive statistical results, but many of the candidates expressed strong reservations about the distractions caused by working on a computer. Elsewhere, a major US research study found a small but significant tendency for lower scores in online mathematics tests *(NAEP, 2005)*. It should be borne in mind that conventional mathematics assessment in the US is already more reliant on multiple choice/ short answer questions than in the UK, so the tasks involved in the NAEP study were far less adventurous than some shown here.

The most likely explanation for this is simply the extra distractions and demands of using the technology. While text editing and word processing are generally seen as valuable tools for writing, it is questionable whether the computer is a "natural medium for doing mathematics", at least for the 5-16 age range. The use of calculators is still seen as contentious (and, where allowed, is not well-integrated in to the syllabus) while the main generic mathematical ICT tool, the spreadsheet, is most useful for discrete mathematics and modelling tasks that have scant coverage in current curricula.

However, the changes in performance seem to be minor "across the board" reductions that complicate the equation of on-line tests with paper "equivalents" but do not necessarily undermine the validity of the assessment. A more important long-term concern is to monitor the educational validity of the tasks and their influence on the taught curriculum, an aim for which statistically stable results may not be sufficient.

5.3: Design and trialling

The design of much paper-based assessment – especially high-stakes tests – is typically fairly conservative. Computer-based assessment is tending to add more use of colour and graphics to the mix, as well as more advanced animation and interaction. All of these have the potential to influence the performance of test questions, requiring a greatly increased range of skills from task designers.

The work on the World Class Arena showed that the division of this between experienced "traditional" assessment designers, programmers and graphic designers was fraught with difficulties. It is essential to build up a skills base of designers who – if not master of all these disciplines – can facilitate collaboration between experts.

Until the design of computer-based assessment becomes a well practised art, careful trialling and refinement is essential. It is also important that such trialling:

- Includes small-scale, close observations where the emphasis is on how individual students interact with the materials, rather than the collection of psychometric data. Such observations on the age 6-7 tests produced a wealth of detailed feedback on the design of individual items that was difficult or impossible to infer from the large scale data.
- Involves educational designers and programmers/software designers in the observation process, so that they gain practical insight and experience. There is a

tendency for this work to be “compartmentalised” in larger organisations.

- Recognises that it is extremely difficult to assemble a truly representative sample of data for statistical analysis or scaling without large resources. Although an impressive number of individual cases can be easily assembled, the number of schools is almost equally important. This is an issue with any school trial, but when ICT is added to the mix the tendency for technical problems to affect entire schools, or for affluent schools with good ICT facilities to self-select exacerbates the problem.

5.4: Project organisation and management

The experience of the World Class Tests project shows that it is essential for on-line assessment development projects to recognise that this is still an innovative field, not a commodity market, and does not necessarily fit traditional designer/producer/distributor models.

Projects should ensure that:

- Technological aspects are subject to the same sort of oversight as the pedagogical content
- The project is driven by educational needs, employing technology where beneficial, not a desire to use a particular technology
- Where projects are initiated by governments/school boards they should be integrated with – and able to both influence and be influenced by – curriculum reform and ICT provision programs.
- Specifications should not be unnecessarily fixed before the experienced teams and personnel are in place.
- The time needed to design and implement any new ICT infrastructure should be taken into account
- Where feasible, rapid prototyping and pre-trialling of tasks and concepts should be considered before they are implemented in a robust, scalable form
- Clear lines of communication and collaboration are established between educational designers and ICT designers

Footnote

This paper was originally written in 2007 to support a bid under the title “Computers in Numeracy assessment”. Some minor changes have been made to remove content and terminology specific to that bid.

References

Shepard, 1989: Shepard, L.A. (1989), *Why we need better assessments*, Educational Leadership, (46) 7 pp. 4-9

Steen, 2000: Steen, L. (2000), *The Case for Quantitative Literacy* in Steen, L. (ed), *Mathematics and Democracy* pp. 1-22, National Council on Education and the Disciplines

Burkhardt, Pead, 2003: Burkhardt, H., Pead, D. (2003), *Computer-based assessment - a platform for better tests?* in Richardson, C (ed), *Whither Assessment* ISBN 1 85838 501 6 pp. 133-145, Qualifications and Curriculum Authority

Boston 2004: Boston, K., *Delivering e-assessment - a fair deal for learners*, 2004, http://www.qca.org.uk/6998.html

NAEP, 2005: Sandene, B., Horkay, N., Bennett, R., Allen, N., Brasswell, J., Kaplan, B., (2005), *Online Assessment in Mathematics amd Writing* NCES 2005-457, National Center for Educational Statistics, US Department of Education, Washington DC, USA http://nces.ed.gov

Other sources of reference

The book Whither Assessment containing the *Burkhardt, Pead, 2003* paper includes a number of highly relevant articles covering the field.

A selection of problem solving tasks from the *World Class Arena* can be found at: http://www.nottingham.ac.uk/education/MARS/papers/whither/index.htm

The official *World Class Arena* site is at http://www.worldclassarena.org

A literature review of “E-assessment” by Ridgway, J., McCuscker, S. and Pead, D. for Futurelab can be found at http://www.futurelab.org.uk/research/reviews/10_01.htm