Beyond multiple choice: Do e-assessment and mathematics add up?

Sarah Hughes, Irene Custodio, Ezekiel Sweiry and Rose Clesham

Edexcel

ABSTRACT

This research relates to how item type impacts on performance on paper-based and computer-based mathematics assessments. Prompts for this work include the drive towards external summative on-screen assessments, the use of technology in teaching and learning as well as formative assessment and an increased focus on the assessment of mathematical processes.

Previous research (e.g. Do-Hong & Huynh 2007 and Poggio et al. 2005) suggests that scores from onscreen and paper assessments are comparable, whilst others (e.g. Threlfal et al 2007 and Johnson and Green 2006) argue that despite similar scores, the ways of working mathematically which students use on paper and onscreen are qualitatively different for some item types.

This research project focused on GCSE mathematics items, in particular multi-step items and those involving diagrams and graphs. Equivalent paper-based and on-screen tests were trialled with 1500 school students. Four strands of evidence were collected and analysed including item scores and students' jottings on paper.

Findings indicated that students performed 3% better on paper than onscreen across the whole test and on one third of the items performance was significantly better on the paper version. The gap between paper and computer performance was significant for items types where students annotated diagrams or graphs on the paper version, which they could not do onscreen (even though were provided with jotting paper). Significant differences were more marked as the demand of the items increased, indicating that more able students were more disadvantaged in an onscreen environment.

There are significant implications relating to understanding the relationship between mathematical thinking and assessment, and whether the development and use of onscreen tools can replicate the ways that students work on paper. Consideration also needs to be given to ensure that new and emerging assessment styles and modes can be applied with validity.

1 Background and context

A number of current issues prompted this research:

1. There is a changing landscape of **computer use in today's classrooms** with more learning and assessment incorporating technology. For assessment to be valid it must assess the skills, knowledge and understanding which are intended. This raises the question of what impact working on paper or on screen has on mathematical thinking and whether it is valid to assess using one mode the mathematical thinking which was taught using another mode.

2. **Concerns about the equivalence** of paper tests and e-assessment. In 2003 Martin Ripley, then working at QCA, stated that e-assessment was inevitable for assessing National Curriculum Tests. The subsequent demise of National Curriculum Tests shows that we cannot always predict the educational future, but nonetheless research in the area of e-assessment around that time focussed on how paper tests could be translated onto screen versions whilst maintaining equivalence between modes and not disadvantaging students. The same concerns currently exist around e-assessment used in GCSEs.

3. **Previous research** has tended to find that scored gained by students taking the same mathematics test onscreen and on paper are similar enough to be considered the same. The vast majority of this research has focussed on multiple choice and short objective questions. QCA's consensus statement made in 2009 took as read that multiple-choice questions delivered onscreen and on paper are equivalent. However, a smaller number of research papers show that although the scores are very similar the ways of thinking and working which students use on paper and onscreen are different for some question types (Threlfal et al 2007, Edecxel 2006, Johnson and Green 2006).

4. **The need for a range of question types in tests.** In regulated exams it is stipulated that a range of question types are used, ensuring that no exam can rely heavily on multiple-choice questions. Similarly, there is a move in mathematics towards more focus on processes in mathematics alongside mathematical content. In response to this the debate about equivalence between onscreen and paper assessments has moved beyond multiple-choice question, to questions rewarding partial credit and multi-step questions. At Edexcel mathematics assessments for Functional Skills and GCSEs both use questions which go beyond simple numerical responses, and require that students show their working and evidence of their mathematical strategies and processes. This means that the key issues when assessing mathematics onscreen will relate to giving partial credit and rewarding evidence of mathematical strategies used.

The project reported here focused on GCSE type questions which were trialled on over 1500 students. The key issue addressed relates to how question type, including those in which students show their working, impacts on performance for paper and computer-based assessments. This project focussed on questions which required particular demands associated with answering maths questions, including:

- Making jottings and working;
- Annotating diagrams and given information;
- Using graphs;
- Manipulating visuals and objects onscreen, including ordering items;
- Selecting relevant information from all the information given.

With the growth of onscreen testing there is a need to ensure that students tackling these question types are not disadvantaged by the mode in which they do their assessment. Particularly as process skills in mathematics become more of a focus, and assessments move beyond multiple choice and objective numerical response questions towards the use of more extended response questions in which students need to show evidence of their mathematical thinking.

2 Research questions

The study sought to answer the following questions:

1. Does test mode affect candidate performance in particular types of mathematics questions?
2. Are differences in performance across modes related to candidate ability and difficulty?
3. Are mode differences related to working memory capacity?
4. Are mode differences related to question type?
5. What are the experiences and perceptions of tutors and of students doing the test?

3 Methods

An onscreen and a paper version of each of 24 GCSE Mathematics questions was developed, with the aim that that paper and computer versions were as similar as possible. A large-scale trial of the tests was carried out in fifty-three centres on 1547 students. Each student tackled all twenty-four questions which were split into two sets of twelve questions named A and B. They attempted half of the questions onscreen and the other half on paper.

Figure 1 Trial design

	First test	Second test
Group 1	Onscreen A	Paper B
Group 2	Onscreen B	Paper A
Group 3	Paper B	Onscreen A
Group 4	Paper A	Onscreen B

The test assessed a range of mathematical knowledge and skills covering shape, number, algebra and data handling. It was a non-calculator test. Two-thirds of questions required long mental or multi-step calculations. All students were provided with paper to make jottings.

Four types of data were collected and analysed:

1. Scores were used to calculate item statistics to enable comparisons between the two versions of each question;
2. Questions were rated on each of these characteristics
    - Amount of working required
    - Amount of annotation required
    - Amount of visual presentation
    - Inclusion or use of graphs
    - Density of information

    to enable an analysis of whether question type was related to difference in performance across modes;
3. Jottings were collected and used as evidence of different types of working and strategies;
4. Feedback from students and tutors was collected through questionnaires and interviews.

4 Example of data

In this section the question named ‘Angles’ is used to illustrate the four types of data collected.
The Angles question was a demanding geometry question which required a lot of annotation, was highly visual and contained dense information.

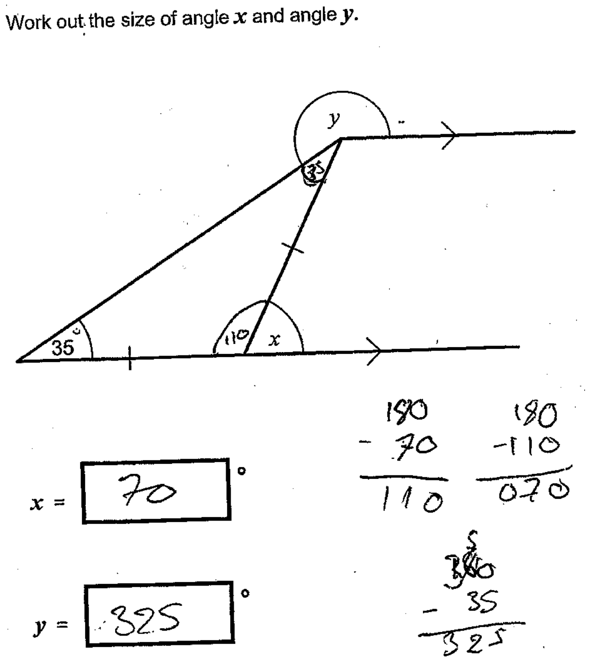
Figure 2 Facility values for Angles question

Item name	Facilities		
	Computer-based test	Paper-based test	Performance difference between modes
Angles_A8A	0.38	0.44	-0.06
Angles_A8B	0.18	0.25	-0.07

The facility values show that part b of the question (angle y) was one of the hardest in the test. The difference in performance between the paper and onscreen versions of the question was 0.06 for angle x and 0.07 for angle y. With students performing better on the paper version than the onscreen version.

Performance on this question was 6% better for angle x and 7% better for angle y on paper than onscreen.

Figure 3 The paper-based version of the angles question



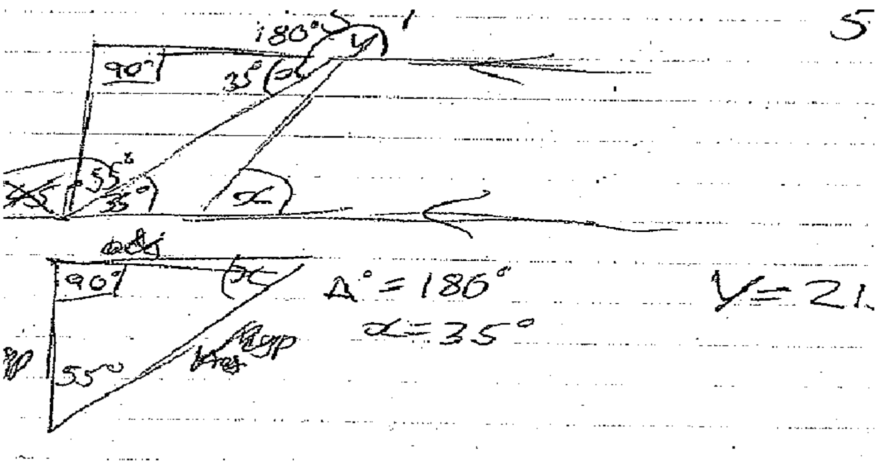
Work out the size of angle x and angle y.

x = 70 °

y = 325 °

Typically, students working on paper annotated the diagram. As shown in this response to the paper-based version of the Angles question.

Figure 4 Student response to the onscreen version of the angles question



Students working onscreen who wanted to annotate the diagram had to redraw the whole diagram, resulting in lost time and accuracy, an example of which is shown in figure 4.

Figure 5 Number of students annotating and showing working for the Angles question

	Students not showing any working or annotation	Students showing Working without annotation	Students annotating and showing working
On paper	24%	20%	56%
Onscreen	60%	26%	14%

Figure 5 shows that most students working onscreen did not use their jotting paper and that most students working on paper annotated the diagram.
Higher ability students did much better when working on paper than onscreen. Showing that on more difficult questions, using the computer disadvantaged the
higher ability students. Performance plotted against ability for the Angles question, angle y.

Figure 6 Item characteristic curve for Angles question 8a

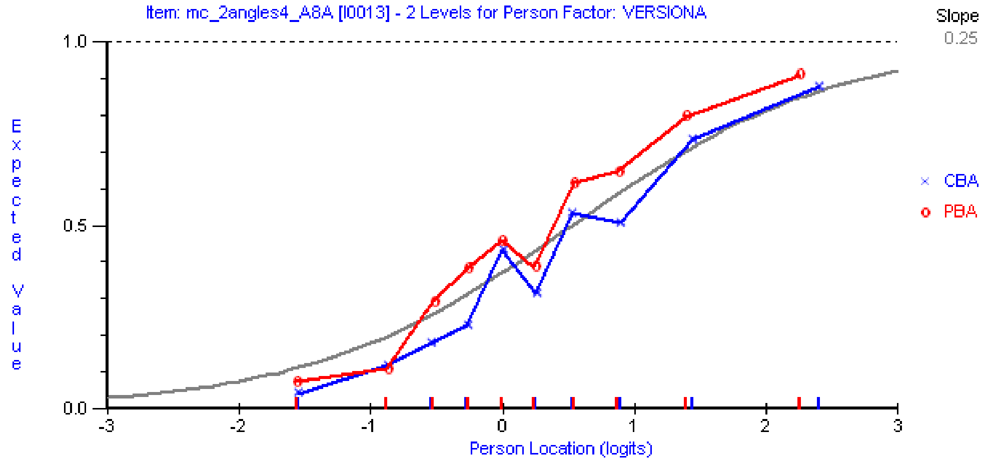


Figure 6 shows that students across the range of ability performed better on the paper-based version than the computer-based version on Test A question 8a.

Figure 7 Item characteristic curve for Test A Q8b

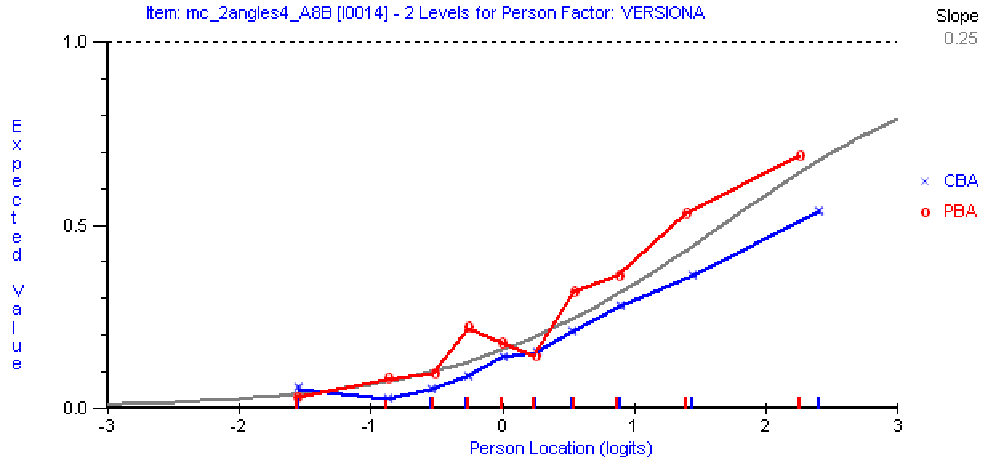


Figure 7 shows that the difference between the two modes is more marked in Question 8b, especially at the higher ability end.

5 Key findings

Key findings are presented in relation to each of the five research questions.

Research question 1: Does test mode affect candidate performance in particular types of mathematics questions?

This question can be split into two:

Research question 1a: To what extent does test mode affect candidate performance?

Test mode affected student performance minimally. Over the total twenty-four questions the average difference in facility value between the modes was 0.03. Showing that on average the scores were 3% higher on paper than onscreen. It is necessary to consider, in real terms, how significant this is in terms of outcomes for our students and equivalence between tests in different modes.

On twelve test items (over nine test questions) there was a statistically significant difference between performance onscreen and on paper, with performance on paper being better than performance on computer. These tended to be questions which required annotation of given diagrams and graphics.

Research question 1b: To what extent does test mode affect candidate performance on particular types of mathematics questions?

Questions were rated on:

- how much working they required;
- how much annotation of given diagrams was required;
- whether questions were presented visually;
- the inclusion of use of graphics;
- the density of information presented.

And these measures were correlated with performance differences between modes to see which of these question characteristics were most closely related to difference in performance between modes.

The amount of working required when answering a question did not impact on any difference in performance between paper and computer versions ($r=0.10$, $p>0.10$). Students all had access to paper for jottings and working out (whether in their test booklet, or as jottings paper alongside their computer) and any numerical or textual working that needed to be done could be carried out.

Those questions where the difference between students' performance on the computer and the paper versions was highest tended to be questions which scored highly on the characteristics 'requiring annotation' ($r=0.55$, $p<0.001$) and 'visual' ($r=0.54$, $p<0.01$). The features 'graphical' ($r=0.38$, $p<0.05$) and 'density of information' ($r=0.30$, $p<0.10$) were also related to performance differences between modes.

Although more students responding on paper showed workings out (the average across all questions was 36% of students) than students working onscreen (24%), this difference was small compared to the difference between the number of students annotating existing diagrams when working onscreen (6%) and when working on paper (29%).

It is likely that students working onscreen were effectively utilising the paper provided for jottings and working. Questionnaire responses showed that students working onscreen used paper for ‘most of the questions’. But there were no resources available for students working onscreen to annotate a diagram. In some cases annotation is the most efficient and appropriate way to tackle a question and the mathematical convention and teaching would be to annotate the diagram as one worked towards an answer. But for students working onscreen this convention was not available to them.

Research question 2: Are differences in performance across modes related to candidate ability and question difficulty?

Item characteristic curves for each item showed where along the ability scale mode differences were concentrated. The sample used in this project ranged from D grade students to B grade students. When we refer to higher ability students, these are the ones around working around GCSE grade B.

Where there were significant differences in performance between modes, these differences tended to be most pronounced around the average and higher ability levels, showing that these students (working at C to B grades) were more likely to perform better on paper than on computer on questions which require working, annotation involve graphics and visuals and include dense information. It looks like the lower ability students are less likely to be able to access the mathematics, and so the impact of mode on their scores is less. Higher ability students, on the other hand, are more impacted by mode because they can access the maths, but their preferred strategies and ways of working may not available to them when working onscreen because these strategies rely more on making jottings and annotating diagrams.

It seems that working onscreen acts a barrier for these higher ability students to demonstrate their mathematical thinking on the harder questions.

Research question 3: Are mode differences related to working memory capacity?

Working memory refers to the ability we have to hold and manipulate information over short periods of time. Questions which would put demands on working memory are multi-step questions containing dense information. Density of information was correlated weakly with performance difference between modes (r=0.30, p<0.10) suggesting that increased demands on working memory maybe related to better performance on paper over computer.

Research question 4: What are the experiences and perceptions of tutors and of students doing the test?

Students in interview talked about their preference for working on paper over computer, and how this was ‘familiar’ and ‘more natural’.

Tutors were concerned that students working on screen would miss out on method marks, would make errors transposing answers from their paper jottings to the computer, would be reluctant to use paper for jottings, and would rely more on doing calculations mentally which could lead to more mistakes. There was also a concern that not being able to annotate a given diagram would disadvantage students who were used to doing particular types of questions (e.g. geometry) by writing or marking on their test paper.

6 Implications

Like much previous research, no significant difference was found between whole test performance onscreen and on paper versions of the same test items. However, there were significant differences in performance on individual items and some methods used by students were qualitatively different across the two modes.

The differences in performance between paper and onscreen questions were higher in questions which required annotation of diagrams and questions which were visual. There was not a significant difference between performance on paper and onscreen questions which required working. These key issues arose:

Paper for jottings

The provision of paper on which to do working out and jottings seems to have brought closer the experiences and performance of students doing calculations onscreen and on paper closer. However, the impact that transcribing paper workings onto an onscreen answer box may have should be investigated.

An onscreen note pad

But a gap still exists between the experience of students working on screen and those working on paper: all students need to be able annotate and work on graphical and diagrammatic questions. The provision of a resource like an onscreen notepad or scribble pad which allows students to write on and around existing diagrams may help close the performance gap between modes for these types of questions. But in designing and evaluating these tools we need to there needs to be an awareness of the impact that they have on students' mathematical thinking.

Performance of higher ability students

On questions where there were significant differences in performance between modes, these differences tended to be most pronounced around the average and higher ability levels. It is possible that this is impacted by a interaction between ability, question type and students' choice of strategy. For questions which require working, annotation and are visual, graphical and/or dense the higher ability students can access the mathematics in these questions, but those working on paper are better able to use their preferred strategies (i.e. annotating and jotting) than those working on screen. It seemed that the more difficult questions were subject to more mode differences. This raises questions about the validity of using e-assessment of mathematics for higher ability students and harder questions.

Validity

It is essential that as assessors we are clear on which constructs are being assessed in each of our qualifications. Without this clarity assessments cannot be fit for purpose or valid. A number of concerns arise when we consider whether e-assessment can be valid or not:

- **The relationship between teaching and learning**. For assessment to be valid students need to be able to demonstrate their mathematical thinking. If this means that they need to be assessed in the same mode as they learnt the mathematics then e-assessment can only be valid for areas of mathematics which are taught using technology. It could be possible that some areas of mathematics may be better suited to one mode than the others.

- **Working to the strengths of paper and of the computer**. Those aspects of maths which are taught using technology (e.g. interactive geometry, data handling and analysis and graphing) may be more appropriately assessed using technology than those areas which are taught and learnt using paper. More understanding is required before it is possible to suggest which areas of mathematics would be better suited to onscreen assessment and which to paper-based assessment.

7 References

Do-Hong Kim & Huynh Huynh (2007), Comparability of Computer and Paper-and-Pencil look back into the future. *Assessment in Education, 10(3), 278-293.*

Poggio, J., Glasnapp, D. R., Yang, X. and Poggio, A. J. (2005), 'A Comparative Evaluation of Score Results from Computerized and Paper & Pencil Mathematics Testing in a Large Scale State Assessment Program, *JTLA .vol 3, no. 6.*

Johnson, M. and Green, S. (2006), online mathematics assessment: the impact of mode on performance and question answering strategies, *JTLA, vol 4, no. 5*

Pommerich (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for

Threlfall, J., Pool, P., Homer and Swinnerton, B. (2007) Implicit aspects of paper and pencil mathematics assessment that come to light through the use of the computer. *Educational Studies in Mathematics. Vol 66 No 3.*