University of Cambridge Local Examinations Syndicate

On-line assessment: the impact of mode on student performance

Martin Johnson and Sylvia Green

A paper to be presented at the British Educational Research Association Annual Conference, Manchester, September 2004.

Disclaimer

The opinions expressed in this paper are those of the authors and are not to be taken as the opinions of the University of Cambridge Local Examinations Syndicate (UCLES) or any of its subsidiaries.

Contact details
Martin Johnson and Sylvia Green
Assessment Directorate
University of Cambridge Local Examinations Syndicate
1 Hills Road
Cambridge
CB1 2EU
Tel: 01223 553843
Email: johnson.m@ucles.org.uk
green.s@ucles.org.uk

On-line assessment: the impact of mode on student performance

Introduction:

The development of new computer technology appears to promise a number of benefits for education and assessment. It has been suggested that the use of computers in the classroom can increase students' intrinsic motivation (Malone, 1981; Lepper, 1988; Guthrie & Richardson, 1995; Schachter, 1999) and lead to greater cognitive gains (BeCTA, 2003). Applying computer technology to educational assessment also promises the opportunity for individualised formative assessment with fewer demands on teachers (Lepper, 1988; Sewell, 1990; Greenwood, Cole, McBride, Morrison, Cowan, & Lee, 2000). At a time when concerns are being raised about the workload burden on teachers, methods aimed at reducing the weight of assessment demands in the classroom are to be welcomed.

Aim:

If computer technology is to be able to fulfil the potential claimed by its supporters, it needs to be seen to at least match the levels of validity and reliability of the paper and pencil assessments that it hopes to replace. Ashton, Schofield and Woodger (2003) argue that contemporary research needs to address a number of issues relating to on-line assessment. They suggest that 'the challenge for on-line assessment is not a technical one, but a pedagogical one. Does the medium matter? Are paper-based questions of the same difficulty as on-line questions?' (2003, p.20). These concerns are not new ones, they echo those of Green, Bock, Humphreys, Linn, and Reckase who stated twenty years ago that 'there is no guarantee that item difficulty is indifferent to mode of presentation' (1984, p.355). This study contributes to that ongoing debate by exploring whether children perform differently according to the mode of assessment, in other words when mathematics questions are presented on computer screen as opposed to when they are presented in traditional paper and pencil form.

A number of studies have already identified a relationship between assessment mode and student performance but, as Bennett (2003) points out, few have investigated this relationship with children of primary or elementary school age. Where this has been done children have generally found questions presented on computer to be more difficult than when presented on paper (Choi & Tinkler, 2002; Coon, McLeod & Thissen, 2002).

This investigation forms part of a wider study which also investigated children's affective responses to working on computers, attempting to gain an insight into the effect of motivational factors (Johnson & Green, 2004). Overall, through the analysis of children's performance and behaviour the study hopes to develop an understanding of how children think when working in two different modes. In so doing, comparing performance and behaviour in the two modes can lead to inferences about how working in the different modes may have affected children's mental processes.

This paper is based on the findings of an initial quantitative analysis of children's performance and the errors that they made when attempting mathematics questions in different modes.

Methodology:

Maths questions were administered to 104 Year 6 (10 and 11-year-old) children in both paper-based and computer-based formats. The children were selected from four primary schools - one large urban school, one small urban school, one large suburban school and one small suburban school. All of the children in participating classes were invited to take part in the study and those gaining parental consent were included.

Two tests, Test A and Test B, each containing 8 questions spanning National Curriculum levels 3 and 4 were constructed (Appendix 1). Each test contained two questions from level 3 and six from level 4. The questions were selected according to a number of criteria. Questions that gave children the opportunity to make their working processes explicit were chosen so that observations could be made about how they approached the problem. To facilitate this children were given a blank piece of paper on which to show their working during each session.

Choosing questions that demonstrated a variety of characteristics was also a consideration, e.g. the response types, the use of tools, the number of 'steps' involved, the level of contextualisation and the type of operation involved. The questions for Test A were matched for difficulty with Test B questions. This was done by matching questions according to National Curriculum criteria and level.

The questions provided ample opportunity for children to show their working and allowed the possibility for further analysis of working methods. When presented online the questions included a replay facility which allowed researchers to view the children's answers as well as other behaviours such as corrections and deletions.

Issues relating to school based access to the internet site hosting the questions and children's ease of navigation through the tests were investigated in a pilot study prior to the main study. As a result of the pilot study two questions which had been initially chosen for the tests were changed. One was considered to be too demanding for children who would be new to Year 6, whilst the other relied on mathematical symbol conventions with which the children felt uncomfortable.

For the main study 104 children took two tests. The test design ensured that at least one week elapsed between children taking their first and second tests. The children were put into four experimental groups. Tests were allocated so that approximately half did Test A first and half did Test B first. This also ensured that approximately half did a paper-based test first and half did a computer-based test first.

	1st test	2nd test	n
Experimental group 1	Test A paper	Test B computer	*27*
Experimental group 2	Test B paper	Test A computer	*26*
Experimental group 3	Test A computer	Test B paper	*26*
Experimental group 4	Test B computer	Test A paper	*25*

Children were randomly assigned to these groups from a sampling frame constructed from lists provided by each of the schools. This was done so that each school had an even number of children and an even gender split within each of the experimental groups, as far as possible.

Children's performance statistics for each question were collected on a database along with gender, teacher assessment level for mathematics, school and group identification data. Data about whether children showed working with their answers was also included. A generic coding frame was built in order to classify types of error (Appendix 2). This framework was compiled after looking at a sample of children's errors made during the tests. Judgements surrounding error classification were moderated during meetings between research team members.

Quantitative findings:

Overall performance

Differences between the teacher assessment levels for children assigned to each group were investigated through an analysis of variance test. This showed that the groups were well

matched (Appendix 3). Data analysis found that there was no statistically significant difference in the overall difficulty of each test. Furthermore it also found that the mode of the test, the order of the test, or whether children answered questions on computer or paper first did not have a statistically significant influence on their results (Appendix 4).

Evidence from facility values for each of the questions appears to suggest that the overall trend was that the paper versions of the questions were marginally easier than the computer versions, although this was not statistically significant (Appendix 5). 11 of the 16 questions were easier on paper than computer. For 3 of these 11 questions the difference was greater than the standard error margin. Only one question had such a difference in favour of the computer version being easier than the paper version. Some differences between modes were small and in a minority of cases the computer version was easier than the paper version. These findings reinforce the need for further investigation to explore how overall test level findings may mask individual question level effects of mode on errors and methods.

Discrimination indices data (Point biserial Correlation) suggested that all of the questions discriminated positively, meaning that they effectively differentiated among children who did well on the overall test and those who did not do well overall (Appendix 6). The question with the lowest discrimination index *(D)* also had a very high facility value *(p)*. Kubiszyn & Borich suggest that this finding is reasonable considering the facility value of the question since 'it can be difficult to obtain discrimination indices above 0.30 when items are easy or difficult' (pp. 126-7, 1990). Questions with a facility value greater than 0.75 are considered to be relatively easy. The data also showed that there was no overall tendency for computer-based questions to discriminate more or less effectively than paper-based questions.

Children's performance was better when they showed their working methods but there was no difference between the frequency of children showing their working methods between modes (Appendix 7). It is important to be cautious here because children were instructed to show their working for both modes since the aim was to collect evidence of their errors and methods for more detailed analysis.

Error Analysis

For both modes computation and mental calculation errors were the most frequent error types. This may not be too surprising since all of the questions involved some degree of computation.

Computation errors were more frequent on computer than on paper. This finding does not appear to be related to children failing to show their working when answering on computer. In many cases children were more likely to show working on their computer test than to show working on their paper test.

Differences in the number of computation errors between modes differed according to the nature of the question. In all instances of questions that demanded subtraction using decomposition children made more computation errors in the computer form of the question than in the paper form.

There was one other error type that appeared to be influenced by mode and the skill demanded by the question. Analysis of errors in the long multiplication questions found that more partitioning errors were made on screen than on paper.

There were relatively few transcription errors but when they were made they were more likely to be on computer. Five children, representing 10% of children in one particular test, had a problem transferring information between screen and page.

Failure to submit an answer to a question was more common on paper than on computer. Interestingly, twice as many boys (n18) than girls (n9) failed to submit an answer to one or more questions in either mode, although this difference was not statistically significant. Boys and girls were both more likely to submit an answer to questions presented on computer, but the difference between modes was more pronounced for boys.

Discussion:

Ashton et al (2003) posed the question 'Does medium matter?', for some of the children in this study the answer appears to be 'yes'. Whilst findings suggest that differences between children's overall performances on paper and computer were not statistically significant, there were enough differences at individual question level to warrant further investigation.

Consistent with the work of others, (Choi & Tinkler, 2002; Coon et al, 2002), this study also suggests that primary aged children generally found questions to be more difficult on computer than on paper. There appear to be a number of possible reasons for this, which have both technical and psychological aspects.

Some children encountered difficulties transferring information from screen to page or vice versa. Although there were relatively few transcription errors overall, when they were made they were more likely to be made when children were attempting computer-based questions. Five children, representing 10% of the children in one particular test, had a problem transferring information between screen and page. This meant that their lack of success should not have been attributed to them having conceptual problems relating to the particular question within which the error was found. This has implications for any system that builds diagnostic profiles based on pupil errors. There is an obvious possibility that there is a potential for misdiagnosis where the cause of error may be due to transcription rather than conceptual problems.

It is interesting to note that transcription difficulties were not found to the same extent when children were making notes for working out and submitting their answers on the same page as the question itself. Most problems occurred when children transferred question information from the screen to their working-out sheet before submitting an answer on screen again. It may be argued that the number of transcription errors is related to the distance that the information needs to be carried, with this distance being greater between the two modes than within the same mode. Computer-based test designers may need to consider incorporating methods that allow children to make notes on screen to minimise problems that children may have when transferring information from one place to another.

There were three questions where children performed significantly better on paper than on computer. For these questions on computer children were less likely to show working on paper and these were the only questions where this was the case. For some reason the children tended not to show written working on these particular questions and this may explain why they were less successful.

It is possible that the question type, the way it is asked and the numbers involved interact with mode to affect willingness to show methods. Simpler questions can be done mentally and it would be expected that mode would have no influence on performance. However for some questions working out on paper would reduce the risk of computation errors, for example when dealing with numbers that 'bridge' tens or hundreds. Children may simply have been 'lazy', preferring to try to do calculations mentally from the screen, whereas on paper it was 'natural' and easier to show working on the page. The distance between the question and the working was less for the paper-based version. Children's error data also appears to support this interpretation. For these three questions children made more combined computational and mental calculation errors when working on computer than on paper. This suggests that a

reluctance to use written methods may have also led children to rely more on mental strategies which contributed to more errors and poorer performance.

Restating the point, if the child thinks the calculation is easy enough they will do it mentally from the screen. If the question is already on paper it is more natural and takes less effort for the child to use written methods to support their thinking. This is where mode may most clearly influence children's strategy choice. If a question is more difficult for a child they will tend to show their working methods in both modes and mode influence will be negligible.

Another interesting finding was that there were a greater number of partitioning problems on screen than on paper. One reason for this may have been because some children perceived numbers in different ways according to mode. One argument is that some children may see numbers presented on screen as fixed entities, interfering with notions that digits may be open to manipulation through the use of flexible strategies. On the other hand, children may be more comfortable with the idea that numbers set down on paper can be played with and their relationships explored in flexible ways. This may be more consistent with children's classroom experience where most of their number exploration will tend to be paper-based.

The suggestion that some children perhaps think differently according to mode may be reinforced by the finding that more children failed to answer questions on paper than on computer. Perhaps this is indicative of how mode may affect attitudes towards working on computer. The data appear to suggest that children, and more specifically boys, were more likely to 'take a chance' about submitting an answer even if they were not sure about whether it was correct. One possible reason for this may be that children may link the activity of answering questions on-screen with other activities commonly associated with computers, such as games, which may promote a philosophy of 'have a go and start again'.

Differences in failing to give answers between modes may also have something to do with possible perceptions that submitting answers on-screen is a less 'personal' activity. When children answer on paper their attempts and errors are made more public, whereas the computer creates a more private workspace where students may be more willing to risk being wrong. When answers are submitted on-line there is no visible trace of evidence relating to past questions which the student may have struggled with, and that they need to confront each time that they look at any subsequent question. This contrasts with the paper versions of each test, which expose children's prior attempts at answers in the public arena occupied by themselves, and potentially their peers and teachers. Having the opportunity to submit answers

in a less public environment may lead children to worry less about the type of answers that they give.

The argument that some children have a different attitude to their answers on computer, being more prepared to 'have a go' and to submit an answer that they haven't fully tested, mirrors findings by Sutherland-Smith (2002) who studied literacy practices and attitudes to computers in Australian primary schools. Sutherland-Smith found that children adopted a 'snatch and grab' philosophy when working on computers. The reasons for this potentially mode-related difference may be influenced by the nature of the activities that children associate with computers outside schools. The connection of computer technology with games is strong and it may be argued that some of the strategies that are successful in a gaming context – such as 'have a go and start again' – may filter into the behaviours of children using computers in other contexts.

This study has raised a number of questions about how mode may have affected the performance of some children. For example, it appears that there may be a relationship between computation errors and mode in certain contexts with questions requiring decomposition or partitioning being apparently more difficult on computer. Furthermore there is a suggestion that the mode of assessment may influence the way that some children may think when answering questions. In order to satisfy concerns about the relative reliability and validity of computer-based and paper-based testing there is scope for more research to probe further any links that may exist between thinking, behaviour and assessment mode.

Bibliography:

Ashton, H.S., Schofield, D.K. & Woodger, S.C. (2003) Piloting Summative Web Assessment in Secondary Education. *Paper presented to 7th International Computer Assisted Assessment conference, Loughborough, July 2003*

Becta (2003) *ImpaCT2: The impact of information and communication technologies on pupil learning and attainment – Full report, March 2003.* Downloaded from http://www.becta.org.uk/research/reports/impact2/

Bennett, R.E. (2003) Online Assessment and the Comparability of Score Meaning. *Paper presented to International Association for Educational Assessment Annual conference, Manchester, October 2003*

Choi, S. W. & Tinkler, T. (2002) Evaluating comparability of paper-and-pencil and computer-based assessment in the K-12 setting. *Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, April 2002*

Coon, C., McLeod, L. & Thissen, D. (2002) *NCCATS update: Comparability results of paper and computer forms of the North Carolina End-of-Grade Tests (RTI Project No. 08486.001).* Raleigh, NC: North Carolina Department of Public Instruction.

Green, B.F., Bock, R., Humphreys, L.G., Linn, R.L. & Reckase, M.D. (1984) Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement* 21 (4) 347-60

Greenwood, L., Cole, U.M., McBride, F.V., Morrison, H., Cowan, P. & Lee, M. (2000) Can the same results be obtained using computer-mediated tests as for paper-based tests for National Curriculum assessment? *Proceedings of the International Conference on Mathematics/Science, Education and Technology*. Vol. 2000 (1) 179-84

Guthrie, L.F. & Richardson, S. (1995) Turned on to language arts: computer literacy in the primary grades. *Educational Leadership* 53 (2) 14-7

Johnson, M. & Green, S. (2004) On-line assessment: the impact of mode on students' strategies, perceptions and behaviours. *Paper presented at British Educational Research Association Annual Conference, Manchester, September 2004*

Kubiszyn, T. & Borich, G. (1990) *Educational testing and measurement: classroom application and practice (3rd Edition)*. Glenview, Illinois: Scott, Foresman & Co.

Lepper, M. (1988) Motivational considerations in the study of instruction. *Cognition and Instruction* 5 289-309

Malone, T. (1981) Toward a theory of intrinsically motivating instruction. *Cognitive Science* 4 333-69

Schacter, J. (1999) *The impact of educational technology on student achievement: what the most current research has to say.* Santa Monica, CA: The Milken Family Foundation

Sewell, D.F. (1990) *New tools for new minds: a cognitive perspective on the use of computers with young children.* Hemel Hempstead: Harvester Wheatsheaf.

Sutherland-Smith, W. (2002) Weaving the literacy web: changes in reading from page to screen. *The Reading Teacher* 55 (7) 664-7

Appendix

Appendix 1

	Test A	Test B
1	There are 472 boys and 18 girls at the cinema. How many children are there altogether?	There are 352 boys and 39 girls at the cinema. How many children are there altogether?
2	Ann scored 554 points in a computer game and Alan scored 538 points. What is the difference in their scores?	Mary scored 546 points in a computer game and Fiona scored 39 points. What is the difference in their scores?
3	At an antique doll fair there are 25 dolls with black hair, 21 dolls with brown hair, and the remaining dolls have fair hair. If there are 90 dolls on display, how many have fair hair?	At an antique doll fair there are 32 dolls with black hair, 18 dolls with brown hair, and the remaining dolls have fair hair. If there are 70 dolls on display, how many have fair hair?
4	Vera went shopping with £70 to spend but only spent £49. She put the rest of the money into her savings account which already had £350 in it. What was the final amount of money in the savings account?	Gavin went shopping with £84 to spend but only spent £43. He put the rest of the money into his savings account which already had £399 in it. What was the final amount of money in the savings account?
5	∀∀ + 5 8 ∀ 1 1	∀∀ + 8 9 ∀ 4 3
6	∀∀ - 2 6 2 9	∀∀ - 4 5 3 6
7	Bob plants 15 rows of turnips in his vegetable garden. There are 25 turnips in each row. How many turnips does he plant?	David plants 15 rows of carrots in his vegetable garden. There are 13 carrots in each row. How many carrots does he plant?
8	What is the perimeter of the following shape? (20cm+20cm+20cm+4cm+8cm+12cm+8cm+4cm)	What is the perimeter of the following shape? (35cm+35cm+35cm+7cm+14cm+21cm+14cm+7cm)

Appendix 2

Error coding types:	
a	non/partial submission (computer only) *failed to give full or partial answer although working shows that child had worked through the answer*
b	transcription error *mistake when transferring information from page to page, screen to page or vice versa*
c	place value error *failed to deal with digits with reference to their place value (there's no obvious 'carrying' leading to computation error)*
d	operation choice *incorrect operation chosen*
e	computation error
f	incomplete *worked through the problem to a point but without reaching a resolution where there is a stop*
g	duplication/over counting/under counting *continued to 'count around' without realising where to finish process*
h	partitioning *confused which numbers to deal with when attempting long multiplication*
i	mental calculation – no working
j	misunderstanding *failing to recognise what the question demands*
k	other
x	no answer

Appendix 3

ANOVA – group ability

ANOVA

paper or computer

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.911	8	.114	.431	.900
Within Groups	25.089	95	.264		
Total	26.000	103			

Appendix 4

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
test	1	200	0.30	0.5856
paper/computer	1	200	0.75	0.3881
test*paper/computer	1	200	1.43	0.2325
quest (test)	14	1400	12.19	<.0001
paper/computer*quest (test)	14	1400	0.80	0.6655

Appendix 5

Facility Value Estimates

test	paper/computer	quest	Estimate	Standard Error	paper-computer
	computer		0.6176	0.02637	
	paper		0.6500	0.02664	0.0324
A	computer		0.6054	0.03747	
A	paper		0.6827	0.03711	0.0773
B	computer		0.6298	0.03711	
B	paper		0.6173	0.03822	-0.0125
A	computer	1	0.8627	0.06521	
A	paper	1	0.8654	0.06458	0.0027
A	computer	2	0.6471	0.06521	
A	paper	2	0.7692	0.06458	*0.1221
A	computer	3	0.6287	0.06521	
A	paper	3	0.7692	0.06458	*0.1405
A	computer	4	0.5490	0.06521	
A	paper	4	0.6923	0.06458	*0.1433
A	computer	5	0.5490	0.06521	
A	paper	5	0.5769	0.06458	0.0279
A	computer	6	0.5294	0.06521	
A	paper	6	0.5769	0.06458	0.0475
A	computer	7	0.4510	0.06521	
A	paper	7	0.5192	0.06458	0.0682
A	computer	8	0.6275	0.06521	
A	paper	8	0.6923	0.06458	0.0648
B	computer	1	0.8462	0.06458	
B	paper	1	0.9184	0.06653	0.0722
B	computer	2	0.6346	0.06458	
B	paper	2	0.5510	0.06653	-0.0836
B	computer	3	0.8846	0.06458	
B	paper	3	0.8571	0.06653	-0.0275
B	computer	4	0.6346	0.06458	
B	paper	4	0.4694	0.06653	*-0.1652
B	computer	5	0.5192	0.06458	
B	paper	5	0.5102	0.06653	-0.0090
B	computer	6	0.5962	0.06458	
B	paper	6	0.5510	0.06653	-0.0452
B	computer	7	0.4231	0.06458	
B	paper	7	0.4898	0.06653	0.0667
B	computer	8	0.5000	0.06458	
B	paper	8	0.5918	0.06653	0.0918

**Difference between paper and computer greater than the standard error margin*

Appendix 6

Point biserial *(D)* and Facility values *(p)*

	A paper		A computer		B paper		B computer	
	(D)	*(p)*	*(D)*	*(p)*	*(D)*	*(p)*	*(D)*	*(p)*
1	0.14	*0.87*	0.07	*0.86*	0.07	*0.92*	0.29	*0.85*
2	0.43	*0.77*	0.57	*0.65*	0.72	*0.55*	0.64	*0.63*
3	0.57	*0.77*	0.50	*0.63*	0.29	*0.86*	0.21	*0.88*
4	0.57	*0.69*	0.64	*0.55*	0.50	*0.47*	0.43	*0.63*
5	0.79	*0.58*	0.65	*0.55*	0.58	*0.51*	0.72	*0.52*
6	0.79	*0.58*	0.72	*0.53*	0.43	*0.55*	0.57	*0.60*
7	0.58	*0.52*	0.79	*0.45*	0.65	*0.49*	0.79	*0.42*
8	0.50	*0.69*	0.36	*0.63*	0.72	*0.59*	0.58	*0.50*

Appendix 7

Total test

ANOVA

total working

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.591	1	.591	.033	.856
Within Groups	1745.769	98	17.814		
Total	1746.360	99			