J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 4, Number 5 · March 2006

On-Line Mathematics Assessment: The Impact of Mode on Performance and Question Answering Strategies

Martin Johnson & Sylvia Green

www.jtla.org

A publication of the Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College

J·T·L·A

Volume 4, Number 5

On-Line Mathematics Assessment: The Impact of Mode on Performance and Question Answering Strategies

Martin Johnson & Sylvia Green

Editor: Michael Russell
russelmh@bc.edu
Technology and Assessment Study Collaborative
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Copy Editor: Kevon R. Tucker-Seeley
Design: Thomas Hoffmann
Layout: Aimee Levy

JTLA is a free on-line journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.

Preferred citation:

Johnson, M. & Green, S. (2006). On-Line Mathematics Assessment: The Impact of Mode on Performance and Question Answering Strategies. *Journal of Technology, Learning, and Assessment, 4*(5). Available from http://www.jtla.org

J·T·L·A

Abstract:

The transition from paper-based to computer-based assessment raises a number of important issues about how mode might affect children's performance and question answering strategies.

In this project 104 eleven-year-olds were given two sets of matched mathematics questions, one set on-line and the other on paper. Facility values were analyzed to explore the impact of the mode on performance. Errors were coded and this allowed further investigation of the differences between questions in the different modes. The study also investigated children's affective responses to working on computer, attempting to gain an insight into the effect of motivational factors. This was made possible by observing and interviewing a sub-sample of children.

Findings suggested that although there were no statistically significant differences between overall performances on paper and computer, there were enough differences at the individual question-level to warrant further investigation. Close analysis of the data suggests that it is possible that the question type, the way it is asked, and the numbers involved, might interact with mode to affect students' willingness to show working methods. The findings also suggest that certain types of questions in certain domains might have different impacts according to mode.

The study concludes that there is scope for more research to probe further any links that may exist between children's thinking, behavior and assessment mode in order to satisfy concerns about the relative reliability and validity of computer-based and paper-based testing.

J·T·L·A

On-Line Mathematics Assessment: The Impact of Mode on Performance and Question Answering Strategies

Martin Johnson[1]
Sylvia Green
Cambridge Assessment [2]

Introduction

Modern technology offers a number of potential opportunities for education and assessment. At a system-level, these opportunities might be manifested in a number of forms. It is feasible that technology might reduce the burden of assessment for teachers by mediating the assessment process. Software developed to communicate information about student task performance can assist in teachers' ongoing assessments and inform their future teaching goals. Through efficiently providing data to both learners and teachers, technology can facilitate the interaction of past student learning (represented by task performance data) and future teaching intentions. It is also possible that by facilitating the interaction of teaching and learning, computer-based assessments might enhance the quality of learning through improved formative feedback, a crucial aspect of formative assessment. According to Black and Wiliam (1998), these are important factors that might affect student motivation and learning.

While recognizing these system-level advantages it is important to explore the relationship between assessment mode and the question answering strategies employed by those being assessed. In the literacy field a debate has developed concerning the effects of mode of communication on thinking structures, and this debate may have implications beyond the confines of literacy. Bearne and Kress (2001) use the term *affordances* to describe "what is made possible and facilitated, and what is made difficult or inhibited" by a medium (p.91). Gibson (1979) has argued that part of the success of human evolutionary development might be a consequence of humans' abilities to exploit the affordances of different environments. Building on this, Wiliam (1999) suggests that unconscious cognitive

processes might play a significant role during the everyday decision making activities of individuals. His analysis suggests that a series of standard configurations or frames of thinking inform the way that individuals choose strategies for action, and that these might be based on their past experience. In the context of this study, it is possible that students might perform differently across dissimilar modes because modal affordances might affect their unconscious cognitive processing when involved in computer-based assessment. Moreover, this effect might be different if they were engaged in paper-based assessment.

Literature Review

The transition from paper-based to computer-based assessment raises a number of important issues about how mode might affect performance. If computer technology is to be able to fulfill the potential claimed by its supporters, it needs at least to 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 and they pose the question, "Does the medium matter? Are paper-based questions of the same difficulty as on-line questions?" (p.20). These concerns are not new ones. They echo those of Green, Bock, Humphreys, Linn, and Reckase (1984) who stated over twenty years ago that "there is no guarantee that item difficulty is indifferent to mode of presentation" (p.355).

Performance and "Administrative Factors"

A number of studies have already investigated the relationship between assessment mode and student performance. Pommerich (2004) suggests that the previous literature seems to indicate that mode differences typically result from the extent to which the presentation of the test and the process of taking the test differ across modes, and not to differences in content. One interesting observation alluded to by Bennett (2003), as well as Russell, Goldberg and O'Connor (2003) is that few studies have investigated this relationship with students of primary- or elementary-school-age. Where this has been done, questions presented on the computer were found to be generally more difficult than when presented on paper (Choi & Tinkler, 2002; Coon, McLeod & Thissen, 2002). In a review of related literature, Russell et al., (2003) suggest that a number of factors have emerged that may influence the validity of computer-based tests. They assert that "administration factors, such as transfer of problems from the screen to scratchwork space, lack of scratchwork space, and inability to review and/or skip individual test items, were found to affect [computer-based] test performance significantly" (p.282). The importance of scratchwork space was also highlighted by Lee and Hopkins (1985) who reported

that it was a salient factor in arithmetic test performance. The implications of this for comparisons between computer- and paper-based performances are made clear by Russell et al., (2003) who conclude that "research on some mathematics tests indicates that validity is threatened when students experience difficulty accessing scratch paper in which they perform calculations" (p.288). In light of this, scratchwork use would be a feature for investigation in the present study, more specifically analyzing whether its use is similar across modes when available to students.

Performance and Domain

In a comparative study of mode effects with students aged 12–14, Greenwood et al., (2000) found that any medium effect was partially domain specific. They found modal differences in performance to be most pronounced when questions involved spatial awareness or gross motor skills, suggesting that these questions were harder on the computer than on paper.

Motivation

For some time 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 improved test scores (British Educational Communications and Technology Agency, 2003). A number of studies have also attempted to investigate factors that influence student attitudes toward assessment mode, in particular. Richardson, Baird, Ridgway, Ripley, Shorrocks-Taylor, and Swan (2002) worked with 9- and 13-year-old students who were identified as being gifted and talented by their teachers and found a generally more favorable reaction to answering questions on the computer compared with paper. There was a variety of reasons given for preferences. Most students alluded to having a generally positive attitude towards computers and this affected their stance towards answering computer-based questions. This hints at the possibility that student motivation for computerized tests may be influenced by their experience of the medium beyond an educational context, and that this attitude is different for paper-based tests. This reinforces work done by Levin and Gordon (1989) who suggested that the dominant consideration affecting student attitudes to working on computers was their prior level of computer experience. This may also help to explain a more contemporary finding by Bridgeman, Lennon and Jackenthal (2002) who reported the comparative popularity of computer-based tests over paper-based tests in a study with American high school students.

Richardson et al., (2002) found other reasons that contributed to student preferences for computer-based questions were related to concrete

differences between the questions in the different modes, such as the use of color illustrations. Other reasons were more clearly related to the affordances allowed by the medium. Some students preferred computer-based questions because they involved typing answers rather than writing with pen or pencil and because their revised answers were neater than if they had been erased on paper.

Study Aim

A fundamental concern of this study is that if computer-mediated assessment is to be a valid and reliable alternative to paper-based assessment, then it is important for developers of computer-mediated assessments to be made aware of the effects that this mode may have on student question answering strategies

This study aims to contribute to the debate about assessment mode and the potential effects on successful question completion by exploring two major themes. The first theme that is explored is whether primary-school-aged students perform differently according to the mode of assessment such as when mathematics questions are presented on a computer screen as opposed to when they are presented in traditional paper and pencil form. Gathering students' performance data as they answer matched questions in different modes allows us to explore whether administrative and domain factors might influence modal performance.

The second theme is an investigation into primary-school-aged students' affective behavioral responses to working on computers. In so doing, the study attempts to gain an insight into the effect of motivational factors and an understanding of why the performances of individuals may differ between modes. Through the collection of supplementary data about how students worked in different modes, inferences are made about the potential effect of mode on their cognitive processes. This allows us to investigate whether the issues that Richardson et al., (2002) found to affect the attitudes of gifted and talented 9- and 13-year-olds also have salience for 10- and 11-year-old students across a wider ability range.

Methodology

Test Construction

Each subject area of the National Curriculum for England is divided into eight level descriptions of increasing difficulty. At approximately age 11 students take National Curriculum tests in English, mathematics and science, and these contain questions from each of the levels 3, 4 and 5. These levels represent the range of expected attainment for students of this age.

For this study two tests were constructed from questions taken from an established test already used in parts of the UK. The questions were based on level-descriptions criteria taken from the Mathematics National Curriculum. According to the National Curriculum, a level description "describes the types and range of performance that pupils working at that level should characteristically demonstrate" (Department for Education and Employment, 1999, p.17). Each test contained 10 mathematics questions spanning levels 3, 4, and 5 of the National Curriculum, mirroring the convention of the National Curriculum tests.

The questions for each of the tests were matched for difficulty according to their National Curriculum criteria and level. Each matched question had the same content and contextualizing features, but with the number variables being changed. For example, question seven in Test A read, *"David plants 15 rows of carrots in his vegetable garden. There are 13 carrots in each row. How many carrots does he plant?"* whereas question seven in Test B read, *"Bob plants 15 rows of turnips in his vegetable garden. There are 25 turnips in each row. How many turnips does he plant?"*. Each test contained two questions from level 3, six from level 4 and two from level 5. Since level 4 is the expected attainment level for the majority of students at this age, it was felt appropriate to include more questions pitched at this level. The questions were selected according to a number of criteria. Questions that gave students the opportunity to make their working processes explicit were chosen so that observations could be made about how they approached the problem. This meant questions were chosen that required students to work through a number of steps, or with large enough numbers that would discourage them from using purely mental strategies. This measure would encourage the students to leave written evidence of their strategies, allowing inferences to be made about their thought processes. Choosing questions that demonstrated a variety of characteristics was also a consideration. These characteristics included the response types, the use of tools, the number of "steps" involved, the level of contextualization, and the type of operation involved.

The tests were administered to 104 10- and 11-year old students in both paper-based and computer-based formats. The students attended four different Cambridgeshire primary schools – one large urban school, one small urban school, one large suburban school and one small suburban school. All of the students in participating classes were invited to take part in the study. Almost all students were given parental consent and were included in the final study.

The overall sample size was chosen so that there would be more than 50 students completing Test A on paper and Test B on the computer, or vice versa. In order to control for test and mode ordering effects (e.g., whether taking Test A or Test B first, or whether taking a computer or paper-based

test first made a difference to student performance) the students were put into four experimental groups. Students were randomly assigned to these groups from a sampling frame constructed from lists of permitted students provided by each of the schools. This was done so that each school had an even number of students and an even gender split within each of the experimental groups, as far as possible (Table 1).

In order to check that the groups had a relatively equal distribution of abilities, *Teacher Assessment* data was collected. In England, teachers are statutorily required to report their students' progress annually against National Curriculum levels. One important element of this requirement is the reporting of Teacher Assessed levels of student achievement. These assessments are often informed by a variety of informal observations and more formal standardized assessment tools. For this study an Analysis of Variance test of Teacher Assessment levels was carried out and this verified that the groups were not significantly different in terms of their reported ability levels ($p>.10$). An important assumption underlying this study was that the full ability range was covered, but that this might not necessarily conform to the exact population distribution.

Table 1: Experimental group design to ensure control of test and mode order

	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

Before any of the students used the tests, all school computers underwent technical checks, to ensure that they had the correct software installed and to check that their display configurations (screen resolution and font settings) were acceptable. Immediately prior to test administration, students were asked to access a practice area where they were able to use the software tools (e.g., the on-screen protractor) and practice the question answer submission process. This session also gave the students the opportunity to raise any questions about using the software. In order to explore the administrative factors related to scratchwork space use across modes, students were asked to show their work where possible. To support this, they were provided with a blank sheet of paper, or scratchwork space, when working on the computer. For the paper-based questions, students showed their work alongside the questions. The process of providing scratchwork space to both conditions also enabled the collection of working-method data for later analysis.

Finally, the software design dictated that students could not preview forthcoming questions on the computer until they had submitted an answer to the question that they were currently viewing. They were also unable to revise past submissions. Both of these administrative factors were obviously available to students as they answered paper-based questions. This difference poses an important concern regarding the comparability of the different mode conditions (Pommerich & Burden, 2000), an issue that will be addressed during the discussion.

Data Collection

In order to answer the questions, "Do students perform differently across modes, and if so, why?" a variety of quantitative and qualitative data were collected. Quantitative data about student performance across modes would help to investigate the first question, while an array of qualitative data about behavioral differences across modes could help to investigate the second.

Quantitative Data

Performance

At the first level, varieties of quantitative data were collected. These data were collected for the first eight rather than for all ten questions. The reason for this was due to software considerations. The software that was used only supported a playback facility showing student actions during the question answering process for eight questions. It was felt that quantitative analysis would be more useful if it were reinforced by the qualitative playback data, which permitted a full coded analysis of errors and also facilitated interviews with students about their working processes.

Since the study was not attempting to establish the validity of the tests, comparisons of facility values would allow investigation into whether particular questions were affected by administration mode. Student performance statistics in the form of facility values for each question were gathered on a database along with gender, teacher assessment level for mathematics, and school and group identification data. Data about whether students showed work with their answers were also included.

Errors

Errors were classified using a generic coding frame (Table 2). This framework was compiled after looking at a sample of student errors made during the tests.

Table 2: Error coding frame

Error Coding Types	
non/partial submission (computer only)	*failed to give full or partial answer although work shows that child had worked through the answer*
transcription error	*mistake when transferring information from page to page, screen to page or vice versa*
place value error	*failed to deal with digits with reference to their place value (there's no obvious "carrying" leading to computation error)*
operation choice	*incorrect operation chosen*
computation error	
incomplete	*worked through the problem to a point but without reaching a resolution where there is a stop*
duplication/over counting/ under counting	*continued to "count around" without realizing where to finish process*
partitioning	*confused which numbers to deal with when attempting long multiplication*
mental calculation – no work	
misunderstanding	*failing to recognize what the question demands*
other	
no answer	

Judgments surrounding error classification were moderated during meetings between research team members.

Qualitative data

Strategies

The second level of data gathering supplemented this quantitative performance data with extra information based on an analysis of student working methods. Where students provided written evidence of their working strategies it was possible to isolate instances where any student's strategies differed on matched questions in different modes. This analysis was used to provide another insight into the thought processes of students as they completed the questions.

Perceptions

At the third level, data were gathered through interviews with a sub-sample of students. Two students were selected from each school. The rationale for sub-sample selection was to include an even gender balance, an even experimental group balance, as well as including students with a mixture of Teacher Assessment levels. T-test analysis showed that the

reported ability of the sub-sample was representative of the 104 students in the larger sample. The structure of interviews was designed to enable students to verbalize their working methods. Students were shown the matched questions and asked to describe if there were any differences in the way that they worked out each of the problems.

This process of Stimulated Recall (Bloom, 1953) was facilitated by the use of a replay option in the computer software that allowed students to see their response and any revisions that they had made during the answering process. Students were also shown any work jottings that they may have made while answering questions. An important aspect of the interview process was to ask the students about their preferences and their supporting reasons for preferring particular questions in each mode.

In order to discern patterns of preferences, each of the sub-sampled students was asked to identify their favorite when they were shown each question on Test A with its matched question from Test B (e.g., Test A question 1 vs. Test B question 1). By combining the responses around each question it was possible to discern whether patterns of preferences were test question or mode related. Analyzing preferences involved a quantitative comparison of the proportions of sub-sampled students who preferred each question in each mode, and a qualitative analysis of the reasons why those preferences were held. The quantitative comparison involved matching up the responses of the sampled students who completed Test A on paper with those who completed Test A on the computer. For example, the majority of the sub-sampled students preferred Test B question 7 over Test A question 7, regardless of whether they attempted the question on paper or on the computer. This suggests that this was a test question rather than a modal effect. On the other hand, a majority of the sub-sampled students who attempted Test A on paper preferred Test A question 8, in contrast to a minority of sub-sampled students who preferred Test A question 8 on the computer. This suggests that students might be influenced by mode rather than question instance.

Sub-sampled students' given reasons for preferences were gathered and investigated qualitatively. Responses were disregarded for analysis purposes if the stated preferences were purely based on the relationships between particular numbers involved in the question. For example, a response that suggested a mathematical rather than a modal influence might be where a student preferred Test B question 3 over Test A question 3 because they perceived calculating "70–50" to be easier than calculating "90–46". On the other hand, preferring Test A question 8 over Test B question 8 because it allowed the student to transcribe the numbers on the paper alongside the question would indicate a modal influence. When responses were gathered for each question, particular patterns of repeated comments were investigated.

Behavior

The fourth level of data collection involved observing the behavior of the sub-sample of students as they completed all ten questions in each of the tests. The overall rationale for the observations was to gain an insight into the effect of motivational factors and capture students' affective responses to working on both computer and paper. These observations were facilitated by the use of a structured, pre-designed observation schedule. The observation schedule included a variety of low-inference measures, these being specific, identifiable behaviors based on feedback from a pilot study, for example, "reading aloud" or "referring backwards and forwards."

Analysis of the observation data involved bringing together records of instances where students exhibited particular behaviors in one mode but not the other. In this way any identifiable patterns could be further interrogated. For example, it was possible to gather evidence of particular students craning their neck for angle measuring questions on the computer, allowing direct comparisons with their behaviors in the matched question in the contrasting mode.

Since the observation schedules used a variety of low inference categories *(e.g., Question 9: Student rotating paper? Y/N)*, it was not felt to be important to establish inter-rater reliability levels. Furthermore, other checks on observer subjectivity were in place. Where an observer felt that there was room for interpretation in relation to a student's actions, the interview provided an opportunity for them to confirm or refute such an interpretation. This process helped overcome concerns about subjectivity and interpretation during observations. Furthermore, the use of a number of observers throughout the study was intended to help negate any dominant assumptions that may have underpinned interpretations made by a single observer.

Quantitative Findings

Overall performance (questions 1–8)

Data analysis found that there was no statistically significant difference in the overall difficulty of each test. Furthermore it was determined that the mode of the test, the order of the test, or whether students answered questions on the computer or on paper first, did not have a statistically significant influence on their results.

Evidence from facility values for each of the questions, in the form of least-squared mean estimates derived after an analysis of variance, 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 (Table 3). Eleven of the sixteen questions were easier on paper than the computer. For three of these eleven questions the difference was greater than the standard error margin. Only one question was easier on the computer than on paper where the difference was greater than the standard error margin. 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.

Table 3: Least-squared mean estimates for the analysis of variance results (questions 1–8)

Facility Value Estimates					
Test	**Paper/ Computer**	**Question**	**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

Table 3: Least-squared mean estimates for the analysis of variance results (questions 1–8) (continued)

Facility Value Estimates					
Test	**paper/ computer**	**question**	**Estimate**	**Standard Error**	**paper-computer**
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

* indicates where the difference between paper and computer is greater than the standard error margin

Discrimination indices data (Point Biserial Correlation) suggested that all of the questions discriminated positively, meaning that they effectively differentiated among students who did well on the overall test and those who did not do well overall. The data also showed that there were no overall tendencies for computer-based questions to discriminate more or less effectively than paper-based questions (Table 4).

There appeared to be some mode-related differences regarding whether students showed their working method with their answers. In nine of the sixteen question instances more students showed their working method for the computer version of the question than for the paper version. Interestingly, this case was only reversed in the case of four question

instances. For three question instances (Test B, questions 1, 3 and 7) the same number of students (n=33, 21, and 37, respectively) showed work in both modes.

Table 4: Discrimination *(D)* and Difficulty *(p)* indices

	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
9	0.43	0.74	0.72	0.48	0.64	0.67	0.43	0.74
10	0.64	0.78	0.72	0.50	0.43	0.69	0.57	0.78

Error Analysis (questions 1–8)

For both modes, computation and mental calculation errors (e.g., these being errors where there was no evidence of written working, therefore inferring an incorrect mental calculation) were the most frequent error types. This may not be too surprising since all of the questions involved some degree of computation. Overall, computation errors were more frequent on the computer than on paper.

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 (e.g., 554–538 or 546–39), students made more computation errors in the computer form of the question than in the paper form.

One other error-type appeared to be influenced by mode and the particular 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. This meant that students made more errors when separating out the Tens and Units components of large numbers, and tended to have more problems multiplying the appropriate parts when working on the computer.

There were relatively few transcription errors but when they were made they were more likely to be on the computer. Five students, representing 10% of the students in one particular test, had a problem transferring information between screen and page, suggesting that this issue may need further investigation.

Failure to submit an answer to a question was more common on paper than on the computer. Interestingly, twice as many boys (n=18) than girls (n=9) 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 the computer, but the difference between modes was more pronounced for boys.

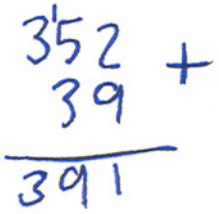
Qualitative Findings

Strategies

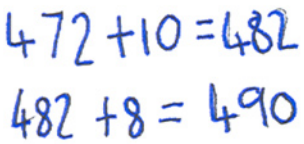
Students' working methods were gathered and analyzed. It was possible to compare strategies for 83 students who showed work for both modes for at least one question. Thirty-nine of these students changed their strategy according to mode. This meant that they chose a different working method when attempting questions that were based on common criteria but where one was attempted on paper and one on the computer. Whether the student got both, one, or neither of the matched questions correct was not considered to be important, since the focus of the study was to capture evidence of process rather than performance.

Although the number of students who changed their method according to matched questions was relatively small it was still possible to discern patterns within some of the questions. For questions that asked students to add two numbers (e.g., 352+39 or 472+18) it was more common for the students to adopt a standard written method when working on the computer. Eight students from the whole sample changed their method for this particular question. Five of these eight chose to use a standard written addition method for the computer versions of the questions while the same number chose to use partitioning strategies (e.g., splitting larger whole numbers into smaller whole numbers prior to operation) when attempting matched questions on paper (Figure 1).

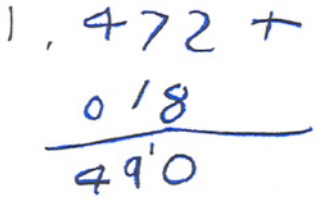
Figure 1:



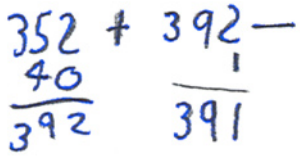
Student 1 computer strategy for 352+39



Student 1 paper strategy for 472+18



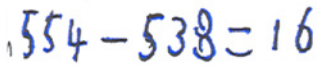
Student 2 computer strategy for 472+18



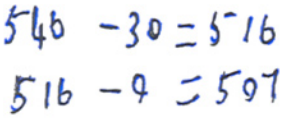
Student 2 paper strategy for 352+39

This tendency to use partitioning on paper rather than on the computer was mirrored in data from questions that asked the students to subtract one number from another (554–538; 546–39). For these questions five of the eleven students from the whole sample who changed their method chose to use partitioning strategies when attempting the questions on paper while only two of this group used this strategy on the computer (Figure 2).

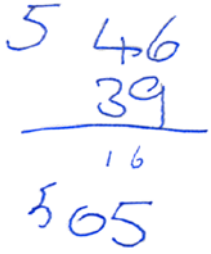
Figure 2:



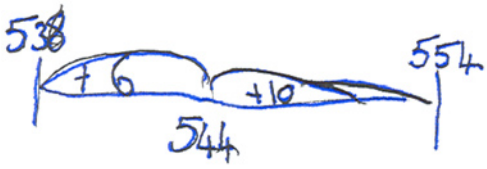
Student 3 computer strategy for 554–538



Student 3 paper strategy for 546–39



Student 4 computer strategy for 546–39

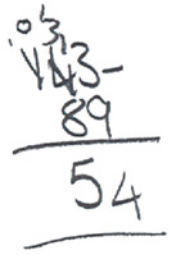


Student 4 paper strategy for 554–538

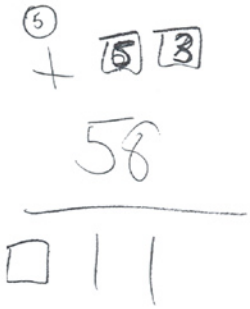
For one of the most difficult questions there appeared to be an interesting mode-related influence on students' strategies. This question was set out as a standard written column addition problem where students were expected to fill in missing digits (☐☐+89 = ☐43; ☐☐+58=☐11). All seven of the students from the whole sample who changed their method for this question adopted an addition-based approach to solving the problem on the computer while three of these students chose to use an inverse subtraction method on paper (see Figure 3).

Figure 3:

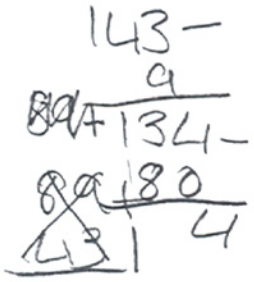
Student 5 computer strategy for ☐☐+58=☐11



Student 5 paper strategy for ☐☐+89=☐43



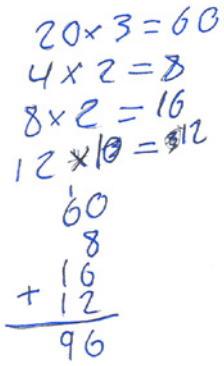
Student 6 computer strategy for ☐☐+58=☐11



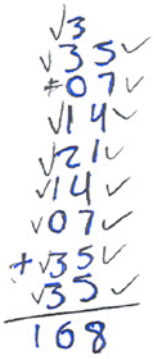
Student 6 paper strategy for☐☐+89=☐43

For the question that asked students to calculate the perimeter of an irregular shape, mode appeared to have an influence on the strategies chosen. Of the nine students from the whole sample who changed their method for this particular question, seven chose a cumulative approach on paper. This meant that they tended to add the measures for each individual side of the shape as they counted around the shape (e.g., 20+4+8+12+8+4...). On the other hand, six of these students chose a combination strategy when working on screen. This meant that they tended to group the numbers relating to matching sides together before combining all of the numbers into a total (e.g., 20+20+20=60, 8+8=16, 4+4=8...) (Figure 4).

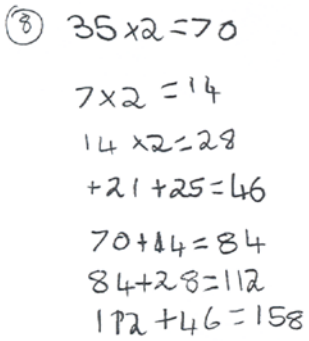
Figure 4:



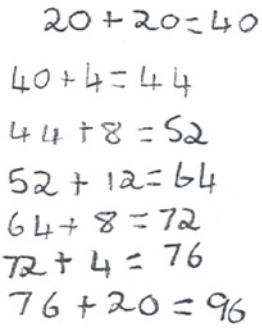
Student 7 computer strategy for calculating the perimeter of a shape



Student 7 paper strategy for calculating the perimeter of a shape



Student 8 computer strategy for calculating the perimeter of a shape



Student 8 paper strategy for calculating the perimeter of a shape

Perceptions

Overall Perceptions

The level of home computer use between the eight sub-sampled students varied. One student had no home computer access while the others generally spent between 30 minutes and one hour per day using a home computer, although the purpose of this use was not probed further.

When asked about their initial feelings about answering questions on the computer, most of the students felt that it was a favorable experience. This mirrors findings by Richardson et al., (2002). Many preferences for particular questions were made based on the specific numbers involved in the questions and were therefore not mode-related. Of the 62 preference judgments expressed, 56.5% were number related and 44.5% were mode related.

When asked about their overall feelings about answering the questions on the computer or on paper, seven of the eight students gave mode-related reasons for their answer. Two liked using computer-based tools and not having to write with pen/pencil. One student felt that he paid more attention to computer-based questions, and the other thought that computer-based questions were less difficult than paper-based questions. Another student felt that computer-based questions contained an element of difficulty that paper-based questions did not. He suggested that computer-based questions restricted his work because he could not write his work down as easily when questions were presented on screen. Despite this he still had a positive attitude toward answering questions on the computer. Two of the eight sampled students felt that the experiences of answering computer-based questions and paper-based questions were similar.

In Favor of the Computer-based Assessment

Five of the eight sampled students felt that computer-based questions were easier than paper-based questions. The greatest generic reason for preferring computer-based questions was the use of the keyboard for word processing rather than writing with a pen/pencil. Most of the students felt that "using keys", "using tools" and "doing less writing" made questions easier.

Other reasons related to question layout being clearer on screen. A number of comments also related to the idea that computer-based questions were more enjoyable than paper-based questions, with one student suggesting that "boring content" could be more fun when presented on screen. The same student also felt that paper-based tests implied time limits, unlike computer-based tests, even though the students were not given a time stipulation for any of the tests. Finally, one student also felt that having to show work out on the question page led to a cluttered and confusing appearance.

In Favor of the Paper-based Assessment

Only two of the eight sampled students felt that computer-based questions were more difficult than paper-based questions. The greatest generic reason for preferring paper-based questions related to not having to transfer attention from page to screen when working out problems. A number of students said they liked their work to be near the question so that they did not have to look away from the problem. These students suggested that switching attention from page to screen to refer to notes contributed to a sense of difficulty, whereas paper-based questions provided a natural space to show their work. The affordance of having space on the page was mentioned as being important for one student who liked to support his work by writing the numbers over the text in the contextualized questions.

The use of the on-screen protractor was also mentioned as a source of difficulty, specifically the manipulation of the protractor around the screen. Angle measuring questions were generally preferred on paper, especially those involving larger angles that required rotation of the protractor. Finally, perimeter calculation questions were generally favored on paper.

A Mixed Picture

The angle measuring questions were problematic. Mode affected student perceptions in a variety of ways in these questions. Half of the sample ($^4/_8$) preferred the paper-based versions of the questions because they felt it was easier to rotate the angle by moving the page without needing to crane their necks in the process. Others felt that it was easier to position the manual protractor compared with the on-screen protractor. Finally, one student felt that the manual protractor was visually clearer than the on-screen version.

Of the other half of the sample who preferred the computer-based versions of the angle questions most comments related to the "fixed" nature of the on-screen protractor. One student felt that the computer protractor stayed more still and "wobbled less" than a manual protractor, while another felt that it was less difficult to position. Another student liked the way that the on-screen protractor could not be placed on the angle "upside down" since its orientation was correct by default. A final comment suggested that the tool introduced an element of "fun" into the question, leading them to pay more attention to the problem.

Student Behavior

In all cases the sampled students completed their paper test more quickly than their computer test. There was only one exception, where one student took an equal length of time for both tests.

"Off task" behaviors were slightly more common on the computer and differed in nature from behaviors observed during paper tests. Three of the eight sampled students were prone to distraction while questions loaded onto their computer but distraction decreased markedly once the students were engaged in answering the questions. On the other hand, inattentiveness during paper tests tended to be caused by distractions elsewhere in the room, such as sudden noise or movement, at any time during the test.

A number of the sampled students exhibited mode-related behaviors when completing the angle measuring questions. Half of the students showed signs of craning their necks while working on the computer but not on paper. Three of the eight also appeared to struggle to read the

on-screen protractor but not the manual protractor. Five of the students adopted a strategy of rotating the paper rather than the protractor when attempting one or both of the angle questions on paper.

Discussion

Recent findings of a study by Poggio, Glasnapp, Yang and Poggio (2005) led them to argue that despite the existence of a few item-level differences across modes, scores from computer-based tests will be equivalent to those obtained from traditional paper-and-pencil tests. They suggest that this is the case "if the computer-based test is constructed in such a way that it reflects the paper-and-pencil version on screen" (p.26). While the findings of the present study are largely in agreement with Poggio et al., (e.g., finding no statistically significant differences between overall performances on paper and on the computer), it can be argued that there were enough differences at the individual question-level to warrant further investigation. As Pommerich (2004) states, "In evaluating mode effects, it is useful to look not only at comparability at the total score level, but also at the item level, because there can be strong mode effects for individual items that cancel out at the overall score level" (p.4). In response to Ashton et al., (2003) it appears that for some of the students in this study the medium of assessment might matter. Consistent with the work of others, (Choi & Tinkler, 2002; Coon et al., 2002), this study also suggests that primary-school-aged students generally found questions to be more difficult on the computer than on paper. There appear to be a number of possible reasons for this, which have both technical and psychological aspects.

Before addressing some of these issues it is important to consider one very important, potentially confounding, issue relating to this study. The findings of Russell et al., (2003) outline the importance of "administrative factors" on test validity. For this study there were observed differences in the ways that students navigated through their test questions. Greenwood et al., (2000) have suggested that secondary-aged students tended not to review their work. In this study, observation evidence suggests that while taking the paper-based test, students tended to review and amend their answers when they had the opportunity. Furthermore, some students navigated through their paper tests by previewing forthcoming questions, apparently "weighing up" whether to attempt some questions before others. It was also possible to observe students reviewing past strategies to inform their approach for new questions. Such observations indicate that the students were seeing the test questions in relation to each other, and that mental processes used were not considered redundant after the closure of an individual question. This supports observations by Pommerich

and Burden (2000) who found that students working through math questions in a cross-subject comparability study showed a greater propensity to skip around the questions than they did when working through questions from other subject areas. They found that "the ability to skip around and the ability to go back and review were very important concerns for the examinees participating in the [math] study" (p.24). It seems students in the present study possessed a degree of independence and control on paper that allowed them access to strategies that could facilitate their performance Furthermore, this independence was compromised by software that prohibited students from going back to earlier questions or from viewing forthcoming questions until they had completed the question at hand. An acknowledged methodological problem of this study is that it is difficult to quantify the extent to which the findings are influenced by such differences. On the other hand, the findings arguably suggest some areas for further investigation that might be robust to this issue. Wiliam's (1999) arguments suggest that there might be merit in capturing the first strategies employed by students, since these might give a unique insight on which to base inferences about student's unconscious cognitive processing. Another potentially worthwhile aspect of this study relates to the close analysis of the relationship between strategy choices and question type data, as well as the study of influences on the likelihood of students to show written working methods.

Another issue that might have been minimally affected by this concern was the technical issue of transcription. Some students 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 found when children were attempting computer-based questions. Five students, representing approximately 10% of the students in one particular testing group, 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, and this raises concerns about validity. This is an area that could benefit from further research, possibly investigating whether students have similar transcription issues if they are asked to use separate scratchwork space when working on paper, rather than using their question paper as a scratchwork resource.

It is interesting to note that transcription difficulties were not found to the same extent when students were making notes for their work and submitting their answers on the paper. Most problems occurred when

students transferred question information from the screen to their scratch paper before submitting an answer on screen again. It may be argued that the number of transcription errors is related to the physical distance that the information needs to be carried during the processing of the problem. This distance might be greater between the two modes than within the same mode. Answering on-screen might require that the question is read on-screen, details held in memory as attention shifts to paper to allow working to be transcribed on paper, then these details are held in memory while attention shifts back to the screen and then the answer is typed into the answer space. This might be contrasted with answering on paper where the question is read, working is transcribed, and the question answered all in close physical proximity to each other. Computer-based test designers may need to consider incorporating methods that allow students to make notes on screen to minimize problems that students may have when transferring information from one place to another.

There were three questions where students performed significantly better on paper than on the computer and here, performances appeared to be influenced by scratch paper For these questions on the computer students were less likely to show their work. It is worth noting that these were three of the only four question instances where this was the case. For some reason the students tended not to show written work 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 (e.g., dealing with numbers that "bridge" tens or hundreds) working out the problem on paper would reduce the risk of computation errors. The distance between the question and the work was less for the paper-based version. Whereas on paper it was "natural" and easier to show work on the page, the extra effort required to support the thinking process on the page while working on screen may have encouraged students to try to do calculations mentally. Student error data also appears to support this interpretation. For these three questions, students made more combined computational and mental calculation errors when working on the computer than on paper This suggests that a reluctance to use written methods may have also led students to rely more on mental strategies that contributed to more errors and poorer performance. Restating the point, if the student thinks the calculation is easy enough he/she will do it mentally from the screen. If the question is already on paper it is more natural, due to familiarity, and takes less effort for the student to use written methods to support his/her thinking. It might be speculated that this is where mode may most clearly influence a student's strategy choice.

If a question is more difficult for students, they tend to show their working methods in both modes and modal influence could be negligible. In some senses this takes further the observations by Russell et al., (2003) that "research on some mathematics tests indicates that validity is threatened when students experience difficulty accessing scratch paper in which they perform calculations" (p.288). In this study, the difficulty in accessing scratch paper was not apparently a physical one but perhaps a mental one, where students had the opportunity to use scratch paper but chose not to use it. This alludes to an interesting relationship between mode and behavior where it may be suggested that strategy choice should not necessarily be expected to be the same across modes.

The findings of this study suggest that mode affected strategy choice for around 37% (n=39) of the students overall. It appeared that students tended to have a more flexible approach to problem solving on paper. Further t-test analyses showed that the while the ability of the students in this group was representative of the larger group, a chi-square test for independence showed that there were significantly more girls (n=26) in this group than boys (n=13) ($p<.001$). When working on the computer students were more influenced by the way that the question was physically presented. This effect was evident in the way that some students approached the matched questions 472+18 and 352+39. For these questions, students tended to approach the problem on the computer by using a standard written addition strategy On the other hand, when working on paper students were less likely to use this formal strategy, instead tending to use informal partitioning strategies. This pattern was mirrored by the way that some students (n=12) approached the matched questions "☐☐+89=☐43" and "☐☐+58=☐11", which were physically laid out in the form of a standard column addition problem. The students who altered their strategies between modes chose to solve this problem on the computer using an addition process, reflecting the manner of its presentation. When attempting the matched problem on paper the most common strategy was inverse subtraction, which might be a more effective approach to dealing with the problem.

Although this study involved relatively small numbers of students, it appears that there was a group of students (n=39) who had a tendency to interpret and behave differently when engaged with screen-based problems compared with paper-based problems, and that these were mostly girls. Furthermore, it appears that some of the students in this group were more likely to apply more flexible strategies to paper-based problems. One suggested reason for this might be that for some students there might be a tendency to view objects presented on screen as being more "fixed" than those presented on paper. If questions presented on screen are taken "at face value" (e.g., problems presented in an "addition" format implying

addition strategies), it is possible that alternate and possibly more effective strategies might be overlooked.

It could be argued that a difference in perception between data presented on screen and on paper may relate to common classroom experience. It is more likely that students will experience mathematical processes involving thinking around problems, manipulating numbers and offering alternative solutions on paper rather than on screen. Furthermore, in the context of the UK it could be inferred that this practice will be more common in the primary-school-years where access to computers is more limited (Department for Education and Skills, 2003) than in secondary-schools. This possible connection between common classroom practice, perceptions of screen-based problems, and strategy use may help to explain the findings of Choi and Tinkler (2002) and Coon et al., (2002) who suggested that primary-aged students found computer-based questions more difficult than paper-based questions.

Mathematical domain also appeared to contribute to the impact of mode on strategy choice, particularly in relation to the shape and space questions. The evidence suggests that students attempting the perimeter calculation questions were more likely to use a cumulative approach when working on paper, and it appears that the affordances of the paper medium promoted this strategy. The data show that students used a more tactile approach to solving the problem on paper, "ticking off" or "dotting" each number around the shape as they accommodated it into their calculation. This approach did not translate into the computer medium where students tended to mentally combine numbers together before calculating the total. The inclusion of this extra "combination" step in the process may have led to a number of students failing to reach an answer on the computer-based perimeter questions. It is interesting that this error type was not found in the matched paper-based perimeter questions. This finding appears to support the work of Greenwood et al., (2000) who found that computer-based spatial awareness questions were more difficult than paper-based questions for secondary-aged students.

The suggestion that some students think differently according to mode may be reinforced by the finding that more students failed to answer questions on paper than on the computer. Perhaps this is indicative of how mode may affect attitudes towards working on the computer. Gallagher, Bridgeman and Cahalan (2000) suggest that it is possible that computer-based testing might create a less threatening environment for some students. The data from this particular study appear to suggest that students, and more specifically boys, were more likely to "take a chance" about submitting an answer even if they were not sure whether it was correct. One possible reason for this may be that students 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 students answer on paper their attempts and errors are made explicit and 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 immediately 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, although this information might be stored elsewhere for teacher analysis at a later time. This contrasts with the paper versions of each test, which expose a student'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 students to worry less about the type of answers that they give, perhaps encouraging them to take risks about which strategies to employ.

The argument that some students have a different attitude towards their answers on the 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 towards computers in Australian primary schools. Sutherland-Smith found that students 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 students 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 behaviors of students using computers in other contexts. This reinforces the view stated by Wiliam (1999), suggesting that unconscious cognitive processes, based on past experience, might play an active role in individuals' decision-making strategies. This argument carries a number of important implications. It could be argued that the influence of a "computer game schema" might influence student perceptions about the true demand of computer-based questions. The logic of this argument implies that a positive disposition to working in the computer medium may lead to a perception that questions presented on the computer may be less demanding than those presented on paper. This could be an important finding since it suggests that students may have a more positive attitude and in turn greater motivation to complete computer-based questions than paper-based questions.

This suggestion might be supported by the interview data from the sample of eight students. The majority of the sampled students felt computer-based questions were easier than paper-based questions. This is an interesting finding when compared with the quantitative performance data since the empirical evidence suggests that computer-based questions were often more difficult than paper-based questions. The notion of ease may be a consequence of both the technical affordances of the computer medium and other perceptual issues connected to students' experiences with computers in the wider environment.

There were a number of physical and technical features of computers that affected student preferences. For most of the students in this study, the concept of "task ease" was related to "doing less writing" (e.g., not having to use pen/pencil for writing). An obvious affordance of the computer medium is the facility to use the keyboard for writing, thereby avoiding manual written activity. As a consequence it appears that computer technology has a built-in advantage over the paper medium since it avoids a crucial area that appears to contribute to students' perceptions of difficulty. This finding supports those of Richardson et al., (2002) who reported a similar reaction from higher ability 9- and 11-year-olds. Other layout features such as the use of color and the combination of colored graphics with supporting text were also felt to make questions easier on the computer. Again, these findings are in agreement with those of Richardson et al., (2003).

It is interesting to note that although students generally preferred answering questions on the computer, there was a group of questions where this trend was reversed. Students preferred shape, space, and measurement questions on paper, supporting the findings of Greenwood et al., (2000).

Another difference between the ways that students behaved according to mode was found within the angle measuring questions. While the computer software only allowed the protractor to be manipulated, observations of paper-based behavior showed a number of students manipulating the paper rather than the protractor. This is another example where the affordances of the technology limited the opportunity for some students to behave on screen as they would on paper. It could be argued that these technical limitations, (e.g., that you measure angle by manipulating a protractor), may discriminate against some students who do not conform to those behaviors. This raises the issue of student experience. Current practice in the UK would tend to suggest that students are more familiar with taking tests on paper than on a computer. This factor cannot be dismissed as it might potentially have influenced the actions of some of the students in this study, despite them having a practice session prior

to their test. As computer-based testing becomes more widespread it is important that students have the opportunity to be as familiar as possible with the experience of test taking on computers so that valid inferences can be made about their ability.

The findings of this study also raise important questions about the merits of transferring questions between modes. It is important that questions are adapted to capitalize on the potential strengths of a mode. In one sense this study attempts to engage with this issue by investigating potential issues that arise when the same students work in different modes. The findings suggest that certain types of questions in certain domains might have different impacts according to mode. Furthermore, this could be because of an interaction between error types, strategy choice, and mode in certain contexts, apparently making some questions more difficult on the computer. 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, behavior and the mode of assessment.

Endnotes

1. Correspondence concerning this article should be addressed to Martin Johnson, Research Division, Cambridge Assessment, 1 Regent Street, Cambridge CB2 1GG, UK; martin.johnson@cambridgeassessment.org.uk
2. Cambridge Assessment is the brand name of the University of Cambridge Local Examinations Syndicate, a not-for-profit department of the University of Cambridge, UK.

References

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

Bearne, E. & Kress, G. (2001). Editorial. *Reading, Literacy and Language, 35*(3), 89–93.

Black, P. & Wiliam, D. (1998). *Inside the Black Box: Raising Standards Through Classroom Assessment.* London: NFER Nelson.

British Educational Communications and Technology Agency [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 at the International Association for Educational Assessment annual conference, Manchester, UK, October 2003.*

Bloom, B. S. (1953). The Thought Process of Students in Discussion. In French, S. J. *Accent on Teaching: experiments in general education.* New York: Harper & Brothers.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2002). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, April 2002.*

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).

Department for Education and Employment (DfEE) (1999). *The National Curriculum: Handbook for primary teachers in England: Key stages 1 and 2.* London: DfEE.

Department for Education and Skills (DfES) (2003). *Information and Communications Technology in Schools in England: 2003.* Available from http://www.dfes.gov.uk/rsgateway/DB/SFR/s000405/index.shtml

Gallagher, A., Bridgeman, B., and Calahan, C. (2000). *The effect of computer-based tests on racial/ethnic, gender and language groups.*, (RR–00–8). Princeton, NJ: Educational Testing Service.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception.* Boston: Houghton Mifflin.

Green, B. F., Darrell 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.

Lee, J. A. & Hopkins, L. (1985). The effects of training on computerized aptitude test performance and anxiety. *Paper presented at the 56th annual meeting of the Eastern Psychological Association, Boston, March 1985.*

Levin, T. & Gordon, C. (1989). Effect of gender and computer experience on attitudes toward computers. *Journal of Computing Research, 5(1),* 69–88.

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.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment 3(6)*. Available from http://www.jtla.org

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment 2(6)*. Available from http://www.jtla.org

Pommerich, M. & Burden, T. (2000). From simulation to application: Examinees react to computerized testing. *Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, April 2000*.

Richardson, M., Baird, J., Ridgway, J., Ripley, M., Shorrocks-Taylor, D., & Swan, M. (2002). Challenging Minds? Students' perceptions of computer-based World Class Tests of problem-solving. *Computers and Human Behavior, 18(6),* 633–49.

Russell, M., Goldberg, A., & O'Connor, K. (2003). Computer-based Testing and Validity: a look back into the future. *Assessment in Education 10(3),* 279–94.

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).

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

Wiliam, D. (1999). The half-second delay: what follows? *Paper presented at the European Conference on Education Research, Lahti, Finland, September 1999*.

Author Biographies

Martin Johnson is a Research Officer in the Assessment Research and Development Division at Cambridge Assessment. Before joining Cambridge Assessment, Martin taught for 10 years in primary schools in the UK. His areas of interest are, amongst other things, the impact of assessment mode on performance and behavior, learners' perceptions of assessment materials, the social implications of assessment, and influences on motivation. Martin has a particular interest in these issues related to vocational and younger learners.

Sylvia Green is Director of Research in the Assessment Research and Development Division at Cambridge Assessment. Current research in the Division is focussed on issues relating to cognitive processes in assessment, comparability and standards, digital assessment and vocational qualifications. Sylvia joined Cambridge Assessment as a research officer in 1994 after having taught in Primary, Secondary and Adult Education. She was appointed Head of the Primary Assessment Unit in 1998, was a senior member of the research team on the Comparability of UK National Test Standards Over Time Project (2003), and from 2003 she was Academic Director of the Formative Assessment Project, developing online materials to support assessment for learning in primary and secondary schools.

J·T·L·A

The Journal of Technology, Learning, and Assessment

www.jtla.org

Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College