Does the mode of delivery affect mathematics examination results?

D. J. Fiddes,* A. A. Korabinski,* G. R. McGuire,* M. A. Youngson* and D. McMillan**
*Heriot-Watt University **Scottish Qualifications Authority
email: g.r.mcguire@hw.ac.uk

At present most examinations are delivered on paper but there is a growing trend in many subjects to deliver some or part of these examinations by computer. It is therefore important to know whether there are any differences in the results obtained by candidates sitting examinations taken by computer compared with those obtained by candidates sitting conventional examinations using pen and paper. The purpose of this article is to describe the outcome of a pilot study designed to investigate possible causes of any differences in results from the use of different modes of delivery in a mathematics examination. One outcome of this study was that the process of translating examination questions into a format required for use on the computer (but keeping this as a pen and paper test) can have a significant effect on examination results. However, the main conclusion is that changing the medium only has no effect on the results in mathematics examinations.

Introduction

There is a growing trend in many subjects to deliver some or part of examinations by computer (Bull and McKenna, 2001; Lawson, 2001). Indeed, in some areas banks of suitable questions are being assembled to allow future examinations in some undergraduate subjects to be taken on paper or by computer (White, 2001; Sims-Williams, 1999). Very little is known, however, about the effects of the medium in testing basic skills. Some studies have been conducted using multiple-choice tests (for example Lee and Weerakoon, 2001, in the area of health education) but mathematics examinations taken on paper and by computer normally consist of questions which the candidate attempts by giving their answer to each question as a number or more generally as a mathematical expression. An up-to-date review of computer-aided assessment in mathematics is presented in a chapter of a recently published book on *Effective Teaching and Learning in Mathematics and its Applications* (Beevers and Paterson, 2002).

Examinations taken by computer have several advantages, not least of which is instantaneous marking and feedback. However, they are different from paper examinations in at least two ways:

- The questions for a computer test require varying amounts of rewording and adjustment in layout compared with the same questions in a paper test in order to write them into a computer assessment package. We shall call this the rewording effect.
- In a computer test, students read questions from the screen and answer these questions by typing in numbers or more general mathematical expressions at specific places (for example an answer box) on the screen. We shall call this the medium effect.

These two effects can create differences in the results of the examination process. If any differences were found between the marks from a paper test and those from its computer version this might be due to either the rewording or the medium effects, or both. The main aim of our project was to devise an experiment to separate these two effects and investigate the significance of each.

Several other factors that might lead to differences between candidate performance in a paper examination and in one using computer delivery include the motivation of the candidate to do well in the examinations, the level of anxiety felt by the candidate when sitting the examinations and the familiarity of the candidate with the assessment software delivering the examinations on computer. The efforts made to deal with these factors through the timing of the experiment and in preparations for the actual tests of the experiment are outlined later in the section on running the experiment.

Setting up the experiment

The Scottish Qualifications Authority (SQA) Higher Mathematics examination has a high uptake in Scotland with virtually all candidates having previously taken examinations from this board. For the experiment, three different test papers were supplied by the SQA, each containing short response questions typical of those found in Paper 1 of Higher Mathematics. The questions used in this pilot project covered most topics in the syllabus with the answers to many questions requiring general mathematical expressions as well as numbers. The marking scheme provided by the SQA was that used in the Higher Mathematics examinations where credit is awarded for each key skill required to be shown by a candidate. In this experiment there was no intention to investigate the most appropriate key skills required in questions, as this would introduce more variables into the experiment. The intention was only to compare current examination practice with its possible computer replacement, and so the marking scheme provided by the SQA was adhered to throughout the experiment and no consideration given as to whether the most appropriate key skills were being examined. While the time allowed for each test was thirty minutes, it was expected that candidates would be able to complete the paper in less time than this so that, in general, time would not be an issue. We use the abbreviation 'P format' for this type of paper test. The questions in the P format tests were then converted into computer test questions as required by the CUE assessment package (Beevers, 2000). Further details of the CUE assessment system and online demonstrations of tests can be obtained at the CALM project Website (CALM Group, 2001). We use the abbreviation 'ICT format' for this type of computer test. In order to separate the medium and rewording

effects, a third type of test was produced, called 'RT format', which is the reverse translation of the computer test and was basically a screen dump of the questions in ICT format. Examples of P format and ICT format questions are shown in Figures 1 and 2.

Paper Question (P format)

1. If $f(x) = 2x - 3$ and $g(x) = 2x^2 - 3$ find an expression for $g(f(x))$. Write your answer in the form $ax^2 + bx + c$.

Figure 1: Example of paper question (P format)

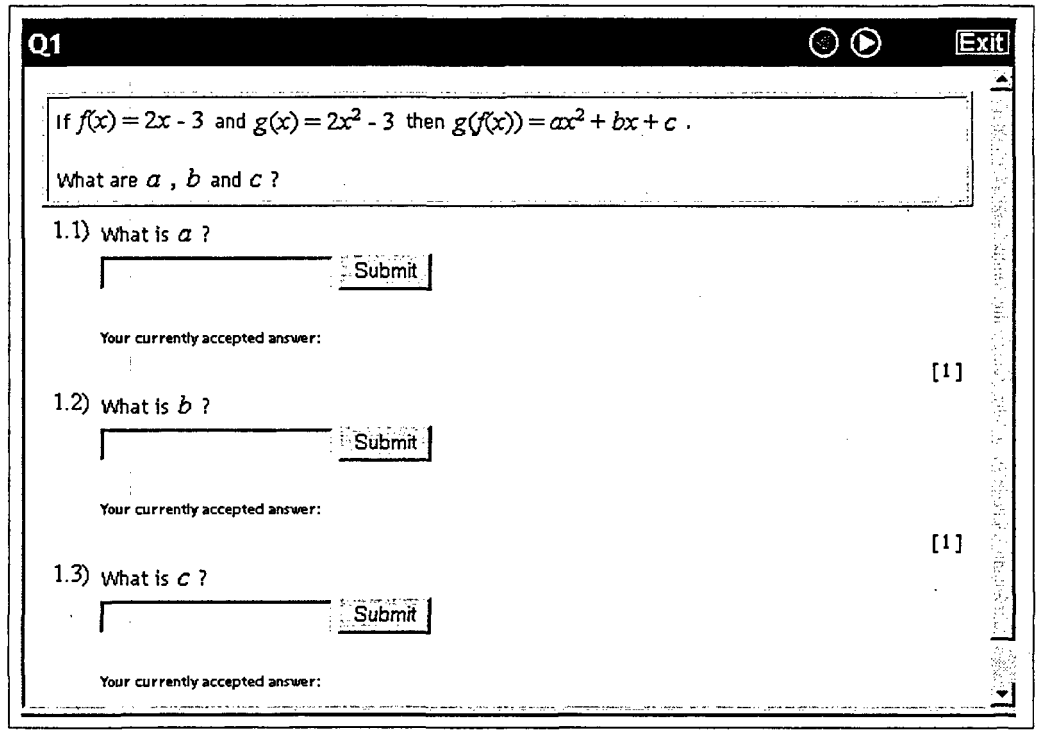

Figure 2: Example of computer question (ICT format)

Although in the examples shown it would have been fairly easy to make the ICT question the same as the P question we wanted to use questions in this project which were reworded so that the rewording effect could be investigated. Since the RT format has exactly the same words in each question and exactly the same place to insert the answer, comparison of the marks between the ICT and RT format (both marked in the same way) should determine the significance of the medium effect.

As both the P and RT format tests are paper tests, there is no change in the medium. However in marking the P format, working is taken into account in giving partial credit for answers that are not correct but in which a candidate has shown some of the skills required to tackle the question. The same could be done with RT format. Space was provided opposite each question for rough working so that in the marking process this working could be taken into account. Therefore the RT paper was marked in two ways. The first was only to mark the answer, which we called RTC marking, since it was exactly how the answers in ICT format tests were marked. The second was called RTW marking since it included giving partial credit for the rough working. Hence comparison of ICT with RTC marks investigates the medium effect and comparison of P with RTW marks investigates the rewording effect.

Running the experiment

Pupils from two schools, 18 from Falkirk High School and 50 from Queensferry High School were invited to participate in the project. All pupils were in their fifth or sixth year of secondary school. There were 40 males and 28 females. Prior to carrying out the experiment, each school was visited, and the pupils were given details of what the project entailed and what would be expected of them. To give them some practice with inputting mathematical answers, a trial ICT test with 5 questions was set up and the pupils were given the opportunity to do this test when help was available to answer their queries about any aspect of what was required to sit the test. The questions in this trial ICT test were in general much easier than Higher standard, but, unlike the questions involved in the project, random parameters were incorporated into the questions. The trial test was available to the pupils from the day of the visit until the day of the experiment so that they could practice as often as they wished beforehand. By introducing random parameters they would get a different test each time they ran the trial. Practice with the trial test would allow pupils to minimize any navigational or inputting difficulties they might have during the computer test and help them gain some familiarity with this new type of test. Some candidates took a lot of advantage of this practice, while others did not.

To enable a paired statistical analysis to be performed on the results of the experiment three groups of pupils were set up at each of the schools in such a way that each group had roughly the same mixture of mathematical ability and gender. Their mathematical ability was estimated from knowledge of their previous SQA examination and Higher preliminary examination results.

The actual tests took place during or just after school hours at each school in April 2001. This was in the small window of opportunity between the time when the pupils had covered enough of the material in the Higher syllabus and before the period when their SQA examinations started. This was a time when the pupils were motivated to do well in

the tests, having been encouraged to regard the tests as good revision for their approaching Higher examination. Each candidate took each of the three tests, one in each of the formats, over a 90-minute period. This was done in such a way that each group sat the three tests in different orders and also that, at any time, no group was sitting the same test or taking a test in the same format as any other group. The computer tests were delivered over the Web with the results being marked and stored at Heriot-Watt University as the pupils sat the tests. During the computer tests, the candidates did not experience either any difficulty in navigating within a test or any delays when submitting answers.

Pupil feedback

The pupils were asked informally by questionnaire what they thought of the tests. Out of a possible 68 pupils, 54 took this opportunity to express their views. The results give some indication about how the pupils felt about the tests they took in the project.

- 47 per cent preferred the paper test, 25 per cent the reverse translation test, 4 per cent the computer test with 24 per cent expressing no preference.
- 4 per cent found the paper test the most stressful, 4 per cent the reverse translation test, 69 per cent the computer test with 23 per cent finding no difference in stress levels between the tests.
- 7 per cent found the computer tests a bit easier than the paper tests, 62 per cent found them a bit harder while 31 per cent found them to be much the same in terms of difficulty.
- 20 per cent thought that using the computer was a better way to be tested in mathematics, 65 per cent thought it was not a better way and 15 per cent felt there was no difference.

Although these results were of some interest, they were not used in the subsequent analysis of the experiment. The pupils who took part in this project had been used to taking tests on paper throughout their school careers. Considering how little experience they had with computer tests, perhaps these views on the computer tests are not surprising.

Marking and analysis

After the pupils sat the tests they were marked as follows. The P format papers were marked by the SQA as though they had been Higher examinations. The RTW examinations were marked at Heriot-Watt using the SQA marking scheme. The ICT tests were marked (automatically) by computer, while the RTC examinations were marked at Heriot-Watt using the same marking scheme that was used by the computer. The statistical analysis involved two separate comparisons. The first was to investigate the rewording effect by comparing P marks with RTW marks. The second was to investigate the medium effect by comparing ICT marks with RTC marks.

Within both schools each of the three tests was taken in each of its three forms by one of the three groups of pupils. With information available on the ability of the pupils, Standard Grade Mathematics results for the Falkirk pupils and Higher Mathematics preliminary marks for the Queensferry pupils, matched pairs were constructed. For

example, one Falkirk pair consisted of two male pupils both with grade 1 in Standard Grade Mathematics, such that one of the pair sat the P version of Test 1 while the other sat the RTW version of the same test. Whereas one Queensferry pair consisted of two female pupils scoring 67 and 68, respectively, in their preliminaries, such that one sat the ICT version of test 3 while the other sat the RTC version of the same test. In this way 14 matched pairs were created from the Falkirk pupils, 11 being matched for ability and gender and 3 for ability but with different genders, and 48 from the Queensferry pupils, all being matched for ability and gender. This resulted in a total of 62 matched pairs. Due to absences of pupils on the day of the tests a very small number of test marks were not used. Statistically a matched pair analysis provides the most efficient use of the data and the sample size of 62 was large enough to validate the analysis without the need for any assumptions such as requiring a normal distribution (McGhee, 1985).

The two statistical analyses were based on the 62 differences in marks for the matched pairs of pupils, either (P – RTW) or (ITC – RTC). In each case the null hypothesis was that the true underlying mean difference was zero against a two-sided alternative hypothesis. A one-sample t-test on these differences was performed using the statistical package Minitab with the following results.

The rewording effect

The mean of the 62 observed differences was –2.3 marks so that RTW marks were greater on average by 2.3 marks in tests, which were marked out of 21 or 22 (see Table 1).

	N	Mean	StDev	SE Mean
P Mark	62	8.484	4.742	0.602
RTW Mark	62	10.774	4.723	0.600
Difference	62	-2.290	4.194	0.533

95% CI for mean difference: (-3.355, -1.225)
T-Test of mean difference = 0 (vs not = 0): T-Value = -4.30 P-Value = 0.0005

Table 1: Paired T-Test and confidence interval for P mark – RTW mark

The observed t-statistic is –4.3 with a probability-value of less than 0.001 to 3 d.p., which is highly significant, giving very strong evidence of a difference between the P marks and the RTW marks.

However, it was noted that the P format papers were marked by SQA while the RTW papers were marked at Heriot-Watt using the SQA marking scheme. Therefore a potential source of variation between the two sets of marks could have been due to the different markers and not just the rewording effect. In order to eliminate any such marker effect and so isolate the rewording effect the P format papers were remarked at Heriot-Watt by the same marker as for the RTW papers, using the same marking scheme. This resulted in some small changes and the data were re-analysed to give the results in Table 2. The set of P marks obtained from remarking is referred to as PMY mark in the output.

	N	Mean	StDev	SE Mean
PMY Mark	62	9.065	4.627	0.588
RTW Mark	62	10.774	4.723	0.600
Difference	62	-1.710	4.079	0.518

95% CI for mean difference: (-2.745, -0.674)
T-Test of mean difference = 0 (vs not = 0): T-Value = -3.30 P-Value = 0.002

Table 2: Paired T-Test and confidence interval for PMY mark – RTW mark

So the mean difference is now –1.7 marks and this can be said to be due to the rewording effect. The probability value is 0.002, which still provides very strong evidence of a difference due to the rewording effect.

These results are illustrated in Figure 3, which gives a histogram of the 62 observed differences together with a 95 per cent confidence interval for the underlying mean and the null hypothesis mean presented below the histogram. The fact that the null hypothesis (Ho) mean is well outside the confidence interval illustrates the very strong evidence of a difference due to the rewording.

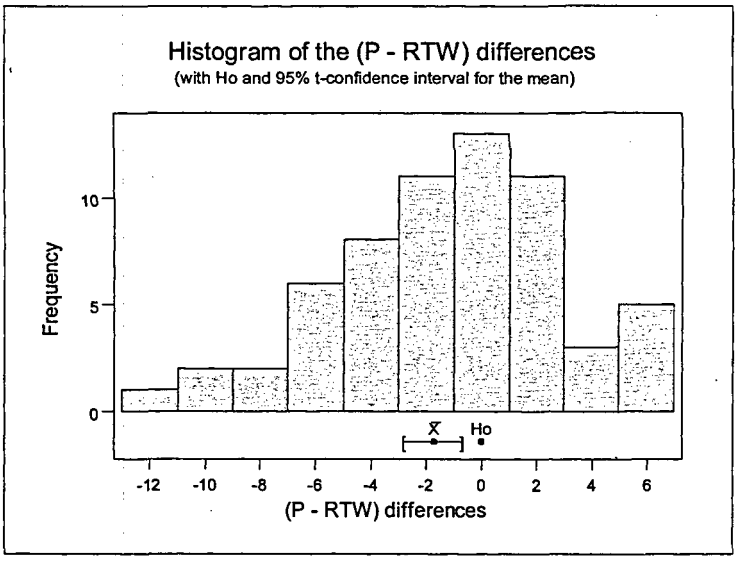

Figure 3

The medium effect

The mean of the 62 observed differences was only –0.2 marks so that RTC marks were greater on average by 0.2 marks in these tests. In this case this is not a significant difference (see Table 3).

The probability value is 0.57 showing that the observed differences could easily have occurred by chance. Again this is illustrated in Figure 4. This time the null hypothesis mean

is well inside the confidence interval, indicating no evidence of a difference due to the medium effect.

	N	Mean	StDev	SE Mean
C Mark	62	7.479	5.654	0.718
RTC Mark	62	7.065	4.534	0.576
Difference	62	0.415	5.738	0.729

95% CI for mean difference: (-1.043, 1.872)
T-Test of mean difference = 0 (vs not = 0):T-Value = 0.57 P-Value = 0.572

Table 3: Paired T-Test and confidence interval for C mark – RTC mark

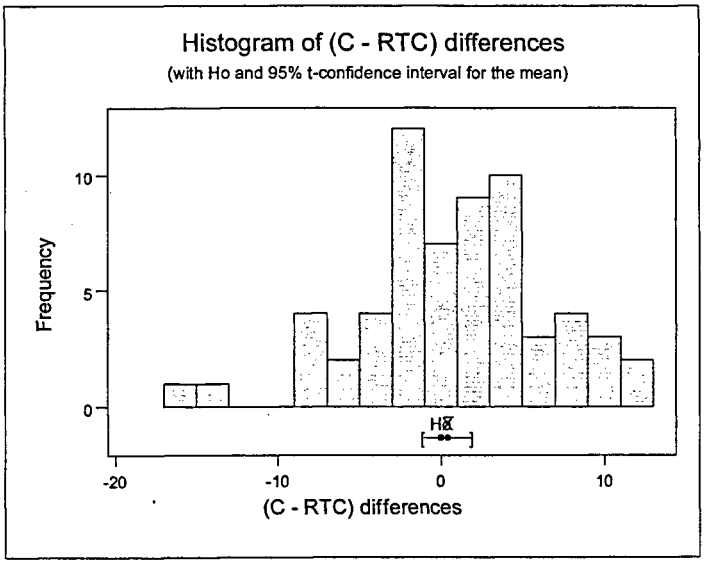

Figure 4

Additional comments

The average mark over all the tests was 39 per cent, with the individual marks varying from 0 per cent to 100 per cent. Although an investigation into a gender effect was not the main purpose of this investigation, a comparison was possible due to the fact that all but three of the 62 pairs were matched for gender. Accordingly the above analyses were repeated for the male pairs and for the female pairs. The results showed that for both the rewording effect and the medium effect the conclusions were the same as above.

Conclusions and directions for further study

The difference in marks between P format and RTW format may be due to the rewording, the formatting (in terms of the number of questions on a page) or the fact that the working for the answers in RTW format was not written in a linear way. In P format, the answers were written with one line following the other making it fairly clear where any error

occurred. If two answers were given in RTW format with one correct and one wrong, it was not as clear which answer the pupils had meant as their answer since, for example, both answers may have appeared side by side. Accordingly, pupils may have been given the benefit of the doubt on occasions. This might be a reason why the RTW marks were higher than P marks. A more detailed study may be required to clarify this issue.

Candidates had little prior experience of computer tests. They were also required to use different mathematical symbols during the computer tests when typing answers into the computer from those with which they were familiar in paper tests. So even with the small amount of practice candidates may have obtained in the trial test beforehand, many would still have some anxiety when sitting the computer test. Despite all of this, the major conclusion drawn from this project is that the medium has no effect on the marks for these tests.

The CUE assessment system has evolved following a number of educational experiments over the last fifteen years. One of its important pedagogic enhancements occurs with the introduction of steps in questions (Beevers, McGuire, Stirling and Wild, 1995). Steps are a possible way of providing partial credit in computer tests (Beevers, Youngson, McGuire, Wild and Fiddes, 1999; Lawson, 2001). There were no steps in any of the questions in the current project. A second investigation is planned into the effect of steps on exam results.

References

Beevers, C. E. (2000), 'Computer aided assessment in mathematics at Heriot-Watt University', *Maths, Stats and OR Newsletter*, 1, 17–19.

Beevers, C. E., McGuire, G. R., Stirling, G. and Wild, D. G. (1995), 'Mathematical ability assessed by computer', *Computers and Education* 25, 3, 123–32.

Beevers, C. E., Youngson, M. A., McGuire, G. R., Wild, D. G. and Fiddes, D. J. (1999), 'Issues of partial credit in mathematical assessment by computer', *ALT-J*, 7, 26–32.

Beevers, C. E. and Paterson, J. S. (2002), 'Assessment in mathematics', in P. Khan and J. Kyle (eds), *Effective Teaching and Learning in Mathematics and its Applications*, London: Kogan Page.

Bull, J. and McKenna, C. (2001), 'Blueprint for computer-assisted assessment'. Available from: *http://www.caacentre/bp*

CALM Group (2001), 'CUE assessment system'. Available from: *http://www.calm.hw.ac.uk/cue.html/*

Lawson, D. (2001), 'Computer-aided assessment in relation to learning outcomes'. Available from: *http://ltsn.mathstore.ac.uk/articles/maths-caa-series/*

Lee, G. and Weerakoon, P. (2001), 'The role of computer-aided assessment in health professional education: a comparison of student performance in computer-based and paper-and-pen multiple-choice tests', *Medical Teacher*, 23, 152–7.

McGhee, J. W. (1985), *Introductory Statistics*, St Paul, MN: West Publishing.

Sims-Williams, J. (1999), 'Open testing with a large databank of multiple choice questions', *Teaching Mathematics and its Applications*, 18, 159–61. (Also *http://www.tal.bris.ac.uk/*).

White, S. (2001), 'Electrical and electronic engineering assessment network'. Available from: *http://www.e3an.ac.uk/*