PARTIAL CREDIT IN MATHEMATICS EXAMS - A COMPARISON OF TRADITIONAL AND CAA EXAMS

G.R. McGuire, M.A. Youngson,
A.A. Korabinski and D. McMillan

Partial Credit in Mathematics Exams - a Comparison of Traditional and CAA Exams.

G.R. McGuire, M.A. Youngson, A.A. Korabinski
School of Mathematical and Computing Sciences
Heriot-Watt University
Edinburgh
EH14 4AS

D. McMillan
Scottish Qualifications Authority
Hanover House,
24 Douglas Street,
Glasgow G2 7NQ

g.r.mcguire@hw.ac.uk

Abstract

With the growing trend in many subjects to deliver at least some part of examinations by computer, it is important to know whether there are any differences in the results obtained by candidates sitting examinations taken by computer compared to those obtained by candidates sitting conventional examinations using pen and paper. The purpose of this paper is to describe a pilot project to compare the traditional type of assessment with assessment done by computer in mathematics examinations and in particular to investigate the role of partial credit in these examinations. In paper based examinations full marks are awarded for a completely correct answer. If, however, a student obtains an incorrect answer but gets some parts of the working correct then in mathematics examinations partial credit is normally awarded. In a computer examination an incorrect answer to a question is normally awarded no marks with no consideration of any partial credit. The mechanism for giving partial credit in the computer examinations of this project was to break the question down into Steps. The project compared results of students taking computer tests in three different formats (either no Steps, compulsory Steps or optional Steps) and the partial credit they would have obtained by taking the corresponding examinations on paper. The tests were at the level of Scottish Higher school examinations and were taken by school students who were about to sit their Higher examinations. This level was chosen as it was high enough to test the students on strategy and mathematical working, while the questions were not too long so that a clearer analysis of the results was possible.

Introduction

While there is a growing trend in many subjects to deliver some or part of these examinations by computer (Bull and McKenna, 2001; Beevers and Paterson, 2002) there has been little analysis which compares results obtained from paper-based examinations with those obtained by ICT examinations. Some studies have been carried out on multiple-choice examinations (for example Lee and Weerakoon, 2001). Other results in this area are contained in Sims-Williams (1999) and White (2001). However most mathematics examinations normally contain other types of question. Typically a question in a paper-based mathematics examination requires the candidate to perform some computations to obtain an answer in the form of a mathematical expression. In ICT examinations it is still expected that the candidate will obtain the answer in this form. Some of the variables which may be involved in changing a paper-based mathematics examination into an ICT examination were discussed in McGuire and Youngson, (2002). To investigate the move from paper-based examinations to ICT examinations it is best to minimise the number of variables that are changed at any stage of the process. An analysis of the role of the medium in ICT examinations is presented in Fiddes et al (2002). The purpose of this paper is to describe an experiment designed to investigate how partial credit may be incorporated into ICT examinations in mathematics. In paper-based examinations a completely correct answer is awarded full marks while a wrong answer with some parts of the working correct may attract partial credit. The need to give partial credit in ICT examinations in mathematics has been recognised by several authors (Beevers et al, 1999; Lawson, 2001: Strickland, 2002).

Setting up the Experiment

Three different test papers, each of thirty minutes duration and containing either 5 or 6 questions of Higher Mathematics standard, were supplied by the Scottish Qualifications Authority (SQA) for the experiment. The questions used in this project broadly covered the whole Higher syllabus and required the students to do several lines of intermediate working before obtaining the final answer. The answers to the questions were, in general, mathematical expressions. The marking scheme provided by the SQA showed how to award credit for each key skill shown by a candidate. The aim of this project was to investigate if this marking process could be replicated in ICT examinations. To avoid introducing more variables into the experiment there was no intention to investigate whether the most appropriate key skills were being examined.

Each test paper was converted into three different ICT examinations each format differing in the amount of help that the candidate was given. The first format of the ICT examinations contained just the original questions with the candidates marked only on their final answers. A candidate who could do some but not all of the question would get no credit for their working if the final answer was incorrect. This format was called the No Steps (NS) format. For the second format, each question was broken down into smaller Steps, each of which the student had to answer correctly to obtain full marks. The Steps corresponded approximately to the method the student would have to go through in order to solve the question. Here the student had to input answers to all the Steps as well as the answer to the original question. This format was

called Compulsory Steps (CS) format. A student who was able to do some of the question but not all of it would be able to demonstrate this knowledge in the CS format. Finally the third format was a hybrid of NS and CS formats whereby in each question the student was originally shown the NS format, but had the option of choosing the CS format if they so wished for that question. This was called the Optional Steps (OS) format. The only proviso in choosing to use the Steps was that if the candidate looked at the Steps then they would have to answer all the Steps correctly to obtain full marks. Although there were no marks deducted for using Steps, their use was not without penalty, because they made a question longer, and there was a time limit for each test. By giving Steps in an ICT examination a strategy for tackling a question might be suggested that would not normally be provided in a paper test. Therefore if a candidate wished to use Steps in OS format, it could be argued that marks should be deducted. However this was not implemented in this experiment in order to make comparisons between OS and CS formats possible. The ICT questions were run using the CUE assessment package. Further details of the CUE assessment system are available at the CALM project website (CALM Group, 2001) and in Paterson, (2002).

Running the Experiment

Pupils from two schools, 16 from Falkirk High School and 26 from Queensferry High School took part in the project. There were 26 males and 16 females. Each school was visited prior to carrying out the experiment when the pupils were given details of what the project entailed and what would be expected of them. In particular they were told about the choice that they would have to make when doing the test with optional Steps. A trial ICT test with 5 questions was set up to give them some practice with inputting mathematical answers and the pupils were given the opportunity to do this test when help was available to answer their queries. These pupils also took part in the experiment described in Fiddes et al (2002). So when they participated in this project, they were familiar with the CUE system and the formats of the questions in the tests. The actual tests took place in late April and early May 2001, just before their SQA examinations. The pupils who took part in this experiment were all due to sit the approaching Higher mathematics examination and were encouraged to think of the tests as good revision for this examination.

Candidates were split into three groups at each school 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. Each group took different tests in such a way that no group sat the same test in the same format as any other group. The candidates were asked to do any rough working in booklets that were collected at the end of the tests. Due to limits on the amount of time which Queensferry High School was able to provide pupils there were able to take tests in only two of the three formats.

Marking

The ICT examinations were marked automatically by computer. In one or two Steps candidates gave alternative correct answers which were not recognised by the computer. In these cases the computer marks were altered to take this

into account. Each format of the examination was also marked in at least one other way using the rough working of the candidates. The rough working was marked to look for any partial credit that may normally have been obtained by the candidates in paper-based examinations.
The NS format with rough working taken into account gave NSW marks and the CS format with rough working taken into account gave CSPC marks. Of all the types of marking, NSW was the one that would most accurately reflect traditional marking of a paper-based examination. In OS format there were two additional markings. The first was to award partial credit in the questions where the students had chosen not to take the steps giving OSPC marks and the second was to give partial credit in all the questions giving OSPC+S marks. The main reason for a difference between CS and CSPC marks for a particular candidate was that if they gave a wrong answer to, let us say, the first part of the question and then subsequently used the right method to the remaining parts, then the computer would give no marks for the remaining parts whereas partial credit would normally accrue in paper-based examinations. The same main reason applied to a difference between OS and OSPC marks for any particular candidate. The differences between NS and NSW marks (and OSPC and OSPC+S marks) could not be assigned in such a simple way as the partial credit was awarded for making variable amounts of progress through each question.
In OS format, 30% of the questions were attempted without the use of Steps. Of these 42% were answered correctly and so 58% were answered incorrectly. Steps were used in each question by 37.5% of the candidates, 54.2% used Steps in at least one question while 8.3% did not use Steps in any question.

Statistical Analysis

Earlier it was noted that the NSW marks are regarded as the closest to marks from a paper-based examination. It is of interest to compare NS with NSW to assess the value of the basic NS format. There were a total of 31 pupils who took an NS test leading to an NS mark and an NSW mark. A paired t-test was performed on these (McGhee, 1985). The data gave an NS mean of 3.8 and an NSW mean of 8.5 with the difference of 4.7 being highly significantly different from zero with a probability-value of less than 0.00005. Therefore there is absolutely no doubt that the basic NS format is not a suitable alternative to paper-based examinations.
In order to compare NSW with OS and CS, matched pairs of pupils were created using the prior knowledge of their abilities. For the Falkirk pupils this was done on the basis of Standard Grade Mathematics grades and for Queensferry pupils on the basis of Higher Mathematics preliminary marks. Gender was also used in the creation of these pairs when possible. For example, one Falkirk pair consisted of two male pupils both with grade 1 in Standard Grade, while one Queensferry pair consisted of a female pupil with preliminary mark 91 and a male pupil with preliminary mark 92. In each pair one pupil sat the NS version of a particular test while the other sat the OS version of the same test. Similar pairings were constructed for NS v. CS comparisons and OS v. CS comparisons. A matched pairs t-test was then applied to perform the required comparisons. Each test involved either 19 or 25 matched pairs and so technically an assumption of normality of the

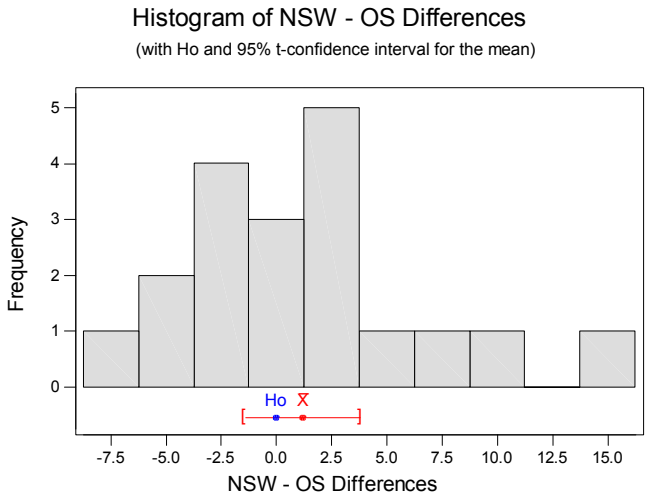
differences (e.g. NS - OS) was necessary for the validity of the analysis. In all cases histograms showed that there were no problems due to a lack of normality.
One of the more relevant comparisons was considered to be NSW v. OS. The results of this comparison are shown below in some detail. This comparison used 19 matched pairs and the resulting mean of the differences (NSW - OS) was 1.2 marks so that the OS marks were less than the NSW marks by 1.2 on average in these tests each of which were marked out of 20 or 21. However this difference of 1.2 was not significant with the observed t-value being only 0.96 and the probability-value being 0.35. The Minitab output is shown below together with a histogram of the 19 observed differences which incorporates a 95% confidence interval for the underlying mean and a point representing the null hypothesis mean (zero).

Paired T-Test and Confidence Interval

```
Paired T for NSW - OS

                 N      Mean     StDev   SE Mean
NSW             19      8.63      5.57      1.28
OS              19      7.45      3.41      0.78
Difference      19      1.18      5.38      1.23

95% CI for mean difference: (-1.41, 3.78)
T-Test of mean difference = 0 (vs not = 0): T-Value = 0.96  P-Value = 0.350
```



In conclusion there is no evidence of a difference between the NSW marks and the OS marks.
A comparison of NSW v. CS was also performed in a similar way. This used 25 matched pairs and again showed no evidence of a difference. Here the mean difference (NSW - CS) was -0.62 so that the CS marks were greater than the NSW marks by 0.6 on average. Again this difference is not significant with a probability-value of 0.51.
Further comparisons of NSW v. OSPC and NSW v. CSPC were carried out. These results are now summarised. NSW v. OSPC: mean difference (NSW - OSPC) = 0.37; probability-value 0.77; no evidence of a difference. NSW v. CSPC: mean difference (NSW - CSPC) = -1.64; probability-value 0.098; slight

evidence, just significant at the 10% level, of a difference with CSPC being a little greater than NSW on average.
The next analyses were comparisons of OS v. CS to investigate any difference between optional and compulsory steps. Again the results are summarised. CS v. OS: mean difference (CS - OS) = 1.7; probability-value 0.098; slight evidence, just significant at the 10% level, of a difference with CS being a little greater than OS on average. CSPC v. OSPC: mean difference (CSPC - OSPC) = 1.9; probability-value 0.091; slight evidence, just significant at the 10% level, of a difference with CSPC being a little greater than OSPC on average. Both of these show that there is weak evidence that compulsory Steps may assist candidates by giving them help with strategy. CSPC v. OSPC+S: mean difference (CSPC - OSPC+S) = 0.55; probability-value 0.54; no evidence of a difference. This last one shows that giving partial credit in all questions brings the OS marks back in line with the CS marks.
Finally NS was compared to OS and CS with the following results. NS v. OS: mean difference (NS - OS) = -3.2; probability-value 0.013; quite strong evidence, almost significant at the 1% level, of a difference with OS being greater than NS on average. NS v. CS: mean difference (NS - CS) = -4.8; probability-value less than 0.00005; very strong evidence of a difference with CS being substantially greater than NS on average. These show that the use of either optional or compulsory Steps offer candidates more scope to show their knowledge compared with no Steps.

Conclusions

Not surprisingly, the candidates' marks in tests without Steps were much lower than those in which Steps were available. They were also lower than those marks that would have been awarded in the corresponding paper-based examinations. This means that without Steps the current marking schemes for paper-based examinations cannot, at present, be replicated by the current computer assessment packages. The longer and more sophisticated the question, the greater the problem.
As noted before, the main reason for a difference between CS and CSPC marks for a particular candidate was that if they gave a wrong answer to, let us say, the first part of the question and then subsequently used the right method to the remaining parts, then the computer would give no marks for the remaining parts whereas partial credit would normally accrue in paper-based examinations. The same main reason applied to a difference between OS and OSPC marks for any particular candidate. It would be helpful to have an assessment package that did not penalise a candidate for the same mistake twice in the same question. The notion of "follow through" (Ashton and Beevers 2002) may provide the facility to deal with such situations. There would then have been no need to consider CSPC and OSPC marks.
There was no evidence of a difference in marks from what would be obtained from a paper-based examination or from a corresponding computer examination with Steps, whether optional or compulsory. Only CSPC marks showed slight evidence of a difference from NSW marks. Also, the CS marks showed slight evidence of being greater than the OS marks as candidates might have had more help with strategy. This would tend to suggest that OS (or, if possible, OSPC marks if "follow through" was available) would reflect the NSW marks most accurately. However, even if the marks obtained are

similar, this does not mean that the candidates have shown the same skills. In particular, the use of Steps provides the candidate with the strategy to do a question. This is normally a skill that a paper-based examination seeks to test. There was no mark penalty for using Steps in this project. If a marking scheme for an examination included strategy marks, it would be possible using the CUE system to penalise students who chose to use the Steps so as to reflect their lack of knowledge of strategy. Clearly in such an examination no student could get full marks using CS format, so in this case the OS format would have to be used. This suggests that one way forward would be to run a new experiment comparing NSW marks to OS (or OSPC) marks with mark penalties for using steps. Candidates in such an experiment would have to be warned about the penalties prior to sitting any test. This new experiment could check both whether there is any difference in the marks and whether the learning outcomes have been examined.
It was also of interest to compare the marks in different formats of those candidates who generally performed badly. These candidates, who perhaps did not have a good grasp of strategy, were able to show some knowledge of the subject in a computer examination with Steps which was not shown in any other type of examination. Many teachers have heard students say that if only they knew where to start a question then they could have done it. The use of (Optional) Steps gives a way of achieving this. Perhaps replication of paper-based examination performance on computer is not necessarily the correct goal at present. It may be more appropriate to choose whichever medium is best equipped to test any particular skill or learning outcome.

References

Ashton, H.S. and Beevers, C.E. (2002) *Extending flexibility in an existing on-line assessment system,* Proc. 6th Int. CAA Conf., Loughborough.

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 Khan, P. and Kyle J. (eds) Effective Teaching and Learning in Mathematics and its
Applications, Kogan Page.

Bull, J. and McKenna, C. (2001) *Blueprint for Computer-assisted Assessment*, http://www.caacentre/

CALM Group (2001) *CUE Assessment System*, http://www.calm.hw.ac.uk/cue.html/

Fiddes, D.J., Korabinski, A.A., McGuire, G.R., Youngson, M.A. and McMillan, D. (2002) *Are Mathematics Exam Results affected by the Mode of Delivery?*, Alt-J 10, 61 - 69.

Lawson, D. (2001) *Computer-Aided Assessment in relation to Learning Outcomes*, http://ltsn.mathstore.ac.uk/articles/maths-caa-series/oct2001/index.htm

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, 152-157.

McGhee, J. W. (1985) *Introductory Statistics*, West Publishing.

McGuire G.R. and Youngson, M.A. (2002) *Assessing ICT Assessment in Mathematics* http://ltsn.mathstore.ac.uk/articles/maths-caa-series/mar2002/index.htm

Paterson, J (2002) *The CUE Assessment System* http://ltsn.mathstore.ac.uk/articles/maths-caa-series/apr2002/index.htm

Sims-Williams, J. (1999*) Open Testing with a Large Databank of Multiple Choice Questions*, Teaching Mathematics and its Applications 18, 159-161. (Also http://www.tal.bris.ac.uk/).

Strickland, N (2002) *Alice Interactive Mathematics,* MSOR Connections 2, 27-30.

White, S (2001) *Electrical and Electronic Engineering Assessment Network*, http://www.e3an.ac.uk/