Abstract

PASS-IT [http://www.pass-it.org.uk/] is a project that has brought together the major Scottish educational agencies with a common goal to investigate how best to exploit e-assessment to enhance flexibility, improve attainment and support teaching and learning in school and college education. PASS-IT has experimented with novel approaches to question design and structure, the role of feedback to support learning and the ability of Computer Aided Assessment (CAA) to measure appropriate learning outcomes in a variety of subjects and levels. Central features of PASS-IT have been to investigate the significance of the medium in test delivery following the work of Greenwood et al [1] and Fiddes et al [2] and the evaluations of the use of steps for partial credit Beevers et al [3] and McGuire et al [4].

The major research findings from phase 1 of the project have been communicated by Ashton et al in two recent papers [5,6]. This present article explains some of the findings of phase 2 of the project with an emphasis on the work in Mathematics.

1. Introduction

PASS-IT was set up to investigate issues associated with online assessment in one subject at all levels of the Scottish school and college curriculum (that is Mathematics from Access levels through all the levels up to and including Advanced Higher) and issues associated with application of online assessment across a range of subjects, mainly, but not exclusively, at Higher. Higher and Advanced Higher are respectively equivalent to AS and A level standards in England.

Phase 1 of PASS-IT identified two main areas to study. The first one was whether there was a medium effect. This had previously been studied in mathematics examinations by Greenwood et al [1] and Fiddes et al [2]. In phase 1, this was extended to chemistry and computing examinations and results from these trials are given in Ashton et al [5]. The second area of study was the evaluation of the use of steps for partial credit and the results of these trials are given in Ashton et al [6].

Phase 2 has extended the range and variety of phase 1 through work in Higher French (both Reading and Listening tests), Intermediate 1 English, and a Listening test in Higher Music (with the Higher French Reading affording another study of the medium effect). In addition, tests in Access 2/3 and Intermediate 1/2 Mathematics were prepared with the latter providing another
evaluation of the use of steps for partial credit. Experiments were carried out in a similar manner to those in phase 1.

Phase 2 of PASS-IT has had a very ambitious programme as measured by the number of pilots, subject levels and question types covered with 625 candidates drawn from 18 examination centres. There have been many technical improvements driven by phase 1 research and feedback as phase 2 progressed. In particular, the reporting system has become an important tool for both the research workers and the classroom teachers. In the former case the interested reader is directed to the work of Ashton et al [7] and Ashton and Schofield [8] and for work relating to classroom practice see the CPD report by Bond [9].

Work in PASS-IT moved along the lines predicted by the American educational guru, Bennett [10] who described the impact of technology on assessment as taking place over three stages:

- Traditional questions converted to electronic format in stage 1;
- An increasing use of multi-media to open up new possibilities in stage 2; and
- The full immersion of the learner within a virtual reality regime in stage 3.

PASS-IT started in stage 1 of the Bennett road map and has moved well into his predicted stage 2 by the end of its phase 2. The interested reader is directed to the work described by Thomas and Milligan [11,12] and Thomas et al [13] for a better understanding of future possibilities now within reach. The results of PASS-IT extend the growing evidence of the role of the computer in both formative and summative testing (see Bull and McKenna [14]).

The main topic of this article is to describe the work carried out in phase 2 which concentrates on the work in Mathematics at the upper end of the school curriculum. An interesting part of PASS-IT has been its diversity. In Advanced Higher Mathematics one teacher in a local Edinburgh school used the PASS-IT assessment system to set some practice questions for her class to prepare them for the summative test. This work was seen as part of the teacher's CPD and the resulting report by Watson [15] makes fascinating reading. Watson explains, in a similar method to that described by Ashton and Youngson [16], how to align learning points with steps and key parts of a question to design good electronic questions.

Responding to teacher pressure from focus groups held after the end of phase 1, PASS-IT embarked on an alternative evidence trial in two Edinburgh schools in Higher Mathematics National Assessment Bank (NAB) unit 2. In this trial, pupils took an on-line examination which counted as part of their assessment. In other PASS-IT trials those doing the experiments knew that the results would not count towards their award. So, in the alternative evidence trial, there was a possibility that pupil behaviour towards the examination would be different. The results of this groundbreaking work are given in the next section.

2. The Higher Mathematics Alternative Evidence trial

The alternative evidence trial gave sixty-five pupils in three classes at two schools the opportunity of sitting Higher Mathematics NAB unit 2 online instead of on paper in December 2003/January 2004. In Scottish Highers pupils must show minimum competence in three units (roughly one per term) before they can take the traditional end of course assessment on which their performance is graded. In the present pilot, in each school participating in the trial, the whole class took the test online. This unit had previously been the subject of a pilot in phase 1 in which the marks did not count (see Ashton et al [6]). In that earlier pilot there was an investigation to see if marks given in a paper based examination (including partial credit) could
be replicated by an online version in which the questions contained steps but where a loss of marks resulted from any step usage.

Throughout the PASS-IT project the sub-team of three learning technologists have worked closely together across the organisations in the consortium. In particular, a meticulous record of all the authored electronic questions has been kept. As changes have been made due to evidence from earlier pilots or programming advances have enabled more appropriate questions to be set on the computer. At that point, such advances have been logged. For example, as the alternative evidence trials began in late 2003 a review of the results from the phase 1 Higher Mathematics pilot took place. In the light of evidence some questions were modified either by a slight change of marking scheme or more radically to produce a major change where the technology had improved. To highlight this process an example is displayed in the next sub-section.

2.1 An example
The illustration is question 1 in the Higher Mathematics NAB unit 2 which can be stated as follows:

Show that \((x+1)\) is a factor of \(f(x) = x^3+2x^2-5x-6\) and express \(f(x)\) in fully factorised form.

The notes from before last year’s pilot remark that since this is a ‘Show that’ style of question, an essay type part could be tried in the first key part to allow for the strategy marks. The wording of this essay question read "To show that \(x+a\) is a factor of \(f(x)\) I would:"". It was felt that multiple choice options gave away too much of the strategy of the question. There was no way around this at that time without giving away part of the answer. It was suggested that this might be a good use of the technology of automatic free text marking in order to score this key part. Students must mention 'synthetic division' and 'remainder zero' to gain the strategy marks.

The second key part then required the fully factorised answer to achieve the remaining marks with suitable restrictions to prevent a student merely typing in the original cubic expression. The outcome from the pilots of phase 1 revealed that this was one of the questions that showed no evidence of a difference with its paper equivalent, although the essay part of the question was removed from the analysis. It was also noted from the pilots that the students found the essay style question very off-putting. To be fair, it is a very different approach than the usual style of question. Nor did it help that this was the first question part the students encountered. There were a mixture of answers with some entering their working into the essay box and others input a word explanation. The diversity in the answers with words and mathematical expressions mingled together meant that this was not a suitable question for automatic free text marking technology at this time.

It was then thought that the use of integrated Flash™ using input boxes would be a better alternative first key part with the students asked to fill in the synthetic division table with the appropriate values. This Flash™ alternative was not available for the pilots in 2003 but in view of the findings at the trials it was created to provide for the synthetic division in time for the alternative evidence trials. This required too that the marks be rearranged. Key part 1 (synthetic division) now carries 1.5 marks (instead of 1 mark for the essay style key part) as the students taking this part will have done the working and achieved the coefficients of the quadratic (in the last line of the synthetic division). It was thought that this would balance the question better as a lot of marks were being lost in the 2003 pilot for arithmetical errors in the second key part.

As a final note on this question it may be that the first key part can now become a step leaving a single key part asking for the full factorisation of \(f\). The better students do not need to show
understanding of synthetic division if they can perform the factorisation by a more robust method. Indeed, this had been the suggestion of the Scottish Qualifications Authority (SQA) Moderator when she first encountered the questions in the autumn of 2003. But, when these questions and the students’ performance were formally moderated in July 2004, the placing of a Flash™ element within a step was thought unnecessary. If opinions change again then such an adaptation is technically possible now since the placing of a Flash™ element within a step is a further advance achieved by the PASS-IT technical team.

2.2 Further details of the Higher Mathematics trial

In the light of the results from phase 1, some of the other questions and marking schemes in the online version given in the 2003 pilot were modified when evidence of a difference had occurred between paper and computer versions of the questions. This sought to minimise such observed differences. These modified versions were shown to the examiners at the SQA for comment before the trial took place to try to ensure that pupils sitting an online version had to exhibit evidence of the same skills to pass the unit as those sitting it on paper. At present, on paper, students take one of five versions of the NAB, which are deemed to be of equal standard.

The schools participating in the alternative evidence trial had taken part in the 2003 PASS-IT pilots in mathematics. Hence, the teachers were familiar with the assessment engine and what they had to teach their pupils on how to log on, navigate through the system, how to enter mathematical expressions and so on. Prior to the trial, pupils had practice in similar questions to become familiar with the system, try out the functionality of the system and how to input mathematical expressions. They also had the benefit during practice of seeing whether their answers were correct from feedback to the questions in either Practice or Help mode. Teachers were able to monitor their pupils’ performance (if they wished) and provide help on a mathematical topic if necessary through the reporting system (see the teacher’s report by Bond [9] for details).

Each class taking the alternative evidence trial was split into two groups: since computers in school laboratories are often close together, candidates were seated from each group alternately. Group A took one version of the unit NAB test while group B took a different, but equivalent, version but both were delivered onto computer. Hence, the questions presented to groups A and B were not the same, reducing the potential for cheating. Pupils were given a login for the test they were to sit (which differed from the one they had used for practice) together with a blank paper booklet and encouraged to put all their rough working into it. All pupils in this trial took an online version of the test in which the pupils had the opportunity to use steps in some questions, but at a cost of losing some marks if they did so. The marks for each question, and the marks that they could still gain in this question if they used steps, were shown on screen so that pupils could make an informed judgement on whether to use the steps or not.

In practice, any exam on computer takes longer than the corresponding examination on paper. For example, logging on: depending on the school system this can take a little time if all the pupils try to do this at the same time. This is, clearly, more time consuming than turning over a page of paper. In addition, examinees have to type the answers at the computer, which takes longer than writing them on paper due to familiarity with the two media. Those who chose to take the option of using steps had to type in all the answers to the subparts as well as giving the final answer to maximise their performance. Some of these time pressures will ease in the future as computers become faster and students more proficient within the electronic medium. For now, to minimise the delays at the start of the test it seems that best practice is to produce a link to the assessment website from the school home page. Then, the pupils can access the test more efficiently as it saves typing in a website address. This eliminates misprints. One school chose to have a pre-prepared seating plan. With this provision, it was
possible to label the rough working booklets on the desks with a pupil's name and login for the test. This saves the pupil time in writing out this information and ensures no two candidates from the same group are sitting next to each other.

At the end of the test the rough working booklets were collected. The marks the pupils scored were available through the reporting system. When a pupil had gained enough marks to pass a learning outcome there was no need to review this. However, if they had not passed a learning outcome then the candidate had some of their answers incorrect. In this case, the working they had produced in the rough working booklet was examined. This was then used in conjunction with the answers they had given online together with the marks already automatically awarded. In some cases, the pupils should have been credited with more marks than they had been awarded automatically by the computer. Generally, there were two reasons for this.

The first reason was that the pupil had typed the correct answer in incorrectly or, even, had forgotten to type it in at all. Such errors can be called input errors. These errors should not be penalised. Input errors were much reduced in the alternative evidence trial compared to the previous experiments reported by Ashton et al [6]. This was due, in part, to this being taken more seriously by pupils since it counted and, partly, due to rewording of questions to reduce the chance of input errors. Moreover, the PASS-iT technical team had created a rendered version of the students’ one-line mathematical expression so that they could see at a glance how the computer was interpreting their string mathematical expression. The input of mathematical expressions had been reported as an issue following the pilots in phase 1. Questionnaire evidence (from all the other phase 2 trials) seems to indicate that this problem has been much improved during the pilots of phase 2. The total amount of marks changed due to input errors is given in the next section.

The second reason was that the pupil could have obtained the wrong answer. Even wrong answers in mathematics have, sometimes, some correct working and marks are awarded for this correct working. These are called partial credit marks. Steps were introduced as a way of awarding partial credit but reviewing the rough working booklets gave a way of checking whether this was successful. The total amount of marks changed due to partial credit marks is also given in the next section.

Marks were amended to allow for both these types of errors and the reasons for any amendment were recorded. This now formed the evidence to determine whether candidates passed or failed learning outcomes. In the current trial, this was done in the form of a spreadsheet but could be done online in future with a teacher submitting the amended marks and reasons directly to SQA. For the purposes of research into the comparability of marks between the online version and paper version all input error and partial credit changes were recorded, including those for pupils who had already passed a learning outcome. This would not have to be done outside this research project.

Even after amendment of marks, some candidates failed one or more of the learning outcomes of the unit. Details of the number of failures are given in the following section, along with the marks amended for input errors and partial credit. Failure in units also happens in the traditional paper format and, in these cases, candidates are allowed one further attempt to pass each unit that they have failed by sitting the outcomes they did not pass in an alternative version of the NAB. This practice was followed in the online tests as well with nine re-sit candidates. Performance in the online re-sits was good, but a failure of Edinburgh City Schools’ Network at a critical time of year meant that a small number of pupils from one school did not get the opportunity to re-sit online. These pupils took the re-sit in a traditional paper format.
To ensure that standards have been satisfactorily maintained, each year SQA looks at a sample of responses in NABS that have been teacher marked. Since alternative evidence was presented for the pupils in this trial, these candidates took part in the moderation process for the year 2004. This was to ensure that the moderator could certify that the evidence presented by these pupils warranted their passing the learning outcomes of NAB unit 2. In addition, it gave the opportunity to develop a template for moderation of online assessment when this programme rolls out. The moderation process required access to the raw and amended marks, the reasons for any changes, the pupils’ online answers and their rough working booklets. If a candidate had been required to re-sit a unit, the same information about their performance in the re-sit was provided. For this moderation process paper copies of all of these were provided but for future moderation only the rough working booklets need be there as the other details could be available online. A random sample of 6 candidates was selected from each of the three classes for moderation. These were found to contain satisfactory evidence of attainment of standards required to pass the learning outcomes for that unit.

2.3 Some results from the Higher Mathematics trial
A general linear statistical model, as described by Clarke and Kempson [17], has been employed for the analysis of results in a number of experiments within the PASS-IT project. This model takes account of ability, gender, school and class and a description of the statistical approach can be found in reference [5]. However it has not been possible to make as detailed a statistical comparison in this trial as in the pilots described in [5,6]. It was impractical to ask the teacher to prepare half the class for an online version while the other half were expected to take the examination on paper. Equally, it would have been unfair to prepare all the pupils for an online test and then on the day present them with a paper version as, unlike the other pilots in PASS-IT, in the alternative evidence trials the examination counted as part of their assessment.

Table 1 below gives the number of candidates in each class and the number who failed each of the four learning outcomes at the first attempt. This includes any mark amendments due to input errors and partial credit.

<table>
<thead>
<tr>
<th>Class</th>
<th>Size of class</th>
<th>Number failing outcome 1</th>
<th>Number failing outcome 2</th>
<th>Number failing outcome 3</th>
<th>Number failing outcome 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>A2</td>
<td>25</td>
<td>0</td>
<td>5</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>B1</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1 - Number of students failing outcomes in the three classes

In class A1 at school 1, one candidate failed two outcomes and a second candidate failed one outcome but both passed at their second attempt, also on computer. In class A2 at school 1, five pupils failed three of the outcomes, another three failed two of the outcomes while another five failed just one of the outcomes. Seven of these thirteen candidates in class 2 who had failed then went on to gain passes at a second attempt on computer. Then, the school computer network had problems due to a region wide failure, so the remaining six students took their second NAB attempt on paper. Class B1 at school 2 had a 100% pass rate at the first attempt.

The next table gives the total change in marks due to input errors and partial credit.
### Table 2 - Mark adjustment for input errors and partial credit

To put this in some perspective, there were 43 marks available across the whole test and classes A2 and B1 provide a spread of ability from moderate to excellent students. For both classes A2 and B1 over the whole class there were 1075 marks to be awarded. Changes, for both input errors and partial credit together, ranged from just over 1 percent to just under 10 percent of the overall marks available, including amendments for those who had already passed and calculated for research purposes only.

The only general linear statistical model analysis in this article compares these three classes for any differences in test score (amended for input errors and partial credit) after allowing for gender and ability (as measured by a mark in the traditional end of course diet of examinations in Higher Mathematics in 2004). These give the adjusted means in Table 3 below.

### Table 3 - Comparison of the three classes

**Overall total (maximum score of 43)**

<table>
<thead>
<tr>
<th>Probability-value</th>
<th>0.37</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adj.mean (class A.1)</td>
<td>39.07 (s.e. 0.86)</td>
</tr>
<tr>
<td>Adj.mean (class A.2)</td>
<td>37.49 (s.e. 1.04)</td>
</tr>
<tr>
<td>Adj.mean (class B.1)</td>
<td>39.83 (s.e. 0.95)</td>
</tr>
</tbody>
</table>

**Note**: s.e. denotes standard error or standard deviation.

So, there is no evidence of a difference between these three computer classes after allowing for ability and gender.

Note that the raw marks of the three classes A1, A2 and B1 were respectively 39.8, 33.1 and 42.6. Weaker students may appear to perform better in a NAB test than their Higher Mathematics mark may suggest. This may, in part, be due to the fact that NAB tests are designed to assess minimum competency, and there is less scope for very able students to show their skills. This is demonstrated here by the high mean scores in these tests, marked out of 43.

### 3. Advanced Higher Mathematics

One of the experiments in phase 1 of PASS-IT was to study the role of steps in one of the Advanced Higher unit NABs, see Ashton et al [6]. Following the trials last year focus groups were held to gather teacher opinion. At one of these meetings a couple of teachers at one of the schools indicated that as part of their CPD they would like to work with the PASS-IT team in phase 2 on some identified extension to the work of phase 1. One such piece of work has...
already been described in the alternative evidence trials outlined above and reported by Bond [9]. The other was to issue a laptop to one teacher taking a group of 14 Advanced Higher pupils. This teacher then set a number of questions to cover the NAB learning outcomes, sent them to the Heriot-Watt server where they were displayed for the pupils to use in their homework time. The authoring of the questions was partly facilitated by the User Manual and the Good Practice Guide and followed the techniques for authoring good questions described by Ashton and Youngson [16]. The resulting CPD report written by Watson [15] explains how that teacher undertook the task and her report contains some excellent advice for anyone starting out to create their own questions and is highly recommended to the interested reader.

4. Conclusions

The decision to go ahead with a “for real” experiment in Higher Mathematics started and ended with the teachers. They had suggested such a course of action at the phase 1 focus groups in the Summer of 2003, they took up the challenge with their students in academic year 2003/4, SQA gave the green-light to proceed and the result concluded with one of the participating teachers recommending this approach for the future at a seminar at the Scottish Office in October 2004. The pilot phase in 2003, described in the article by Ashton et al [6], was essential to “fine-tune” all the questions in the NAB. It now seems that this mode of delivery can take its place alongside more traditional measures of student ability following the successful moderation of this work in July 2004. The moderation process was also commended at the same seminar in October 2004, by the SQA chief moderator, herself a teacher.

Coming to terms with the input of the one-line mathematical expression had been a problem for some students in the phase 1 trials. However, two changes between phases 1 and 2 combined to reduce this problem considerably in phase 2 as measured by the questionnaire data. Firstly, an input test was created to give students practice with the typing in of answers and, secondly, a technical advance enabled the one-line mathematical expression to be displayed on screen, as it would be written on paper.

Without doubt the alternative evidence was a great success. Sixty-five students claimed their NAB unit 2 pass in Higher Mathematics with their ability measured mainly by the computer. Most progress has been achieved at the level of Higher Mathematics. It is reasonable to e-assess where e-learning is taking place. With the widescale acceptance of the online courses delivered through SCHOLAR to all Scottish secondary schools (see http://scholar.hw.ac.uk), Scotland stands in a most favourable position to offer its youngsters a qualification online in a range of subjects: Biology, Chemistry, Computing, Mathematics and Physics at Higher and Advanced Higher.

It is likely that further integration of computer assessment with simulation packages, such as the one at http://www.jelsim.org and used to prepare some practical Chemistry simulations as described by Thomas and Milligan [11], will provide a whole new way of testing fundamental understanding in some subjects. Such integration will be powerful in formative assessment to tease out, for example, misconceptions. Indeed, through the integration of computerised simulations and assessment, it may be possible to investigate deep learning and measure higher order skills more effectively than by traditional pen and paper approaches in summative examinations (see Thomas et al [13]). PASS-IT has enabled the research assessment system to enhance its capability in this area and as the second phase draws to a close, it integrates with Flash™ animations, JELSIM simulations and the CHIME animation of molecules.

Randomisation of parameters in formative assessment may be helpful but its use in summative testing has to be used with care to ensure consistency and fairness. This has been investigated during the alternative evidence trials albeit in a limited sense.
Increasingly sophisticated approaches to e-assessment are being developed which raise important issues not only for emerging forms of formative and summative e-assessment, but also for existing traditional forms. Innovative question types that go beyond that possible on paper, such as the inclusion of multimedia, animation, simulation, application of random parameters in Science, Chime molecules in Chemistry or Biology and optional steps all offer valuable opportunities to enhance student learning through formative assessment. But, these approaches also raise challenges to the established culture of assessment, particularly where summative e-assessment is employed. The PASS-IT project has worked to explore these critical issues and has drawn on an extensive body of research and practice in Higher Education. It is clear that automatic testing is here to stay so, it remains important that investigations continue on its validity, reliability and robustness.

The purposes of assessment have been neatly summarised in a diagram to be found on the website at [http://www.ltscotland.org.uk/assess](http://www.ltscotland.org.uk/assess) and reproduced here as Fig 1.

![Fig 1 - Assessment is for Learning diagram](image)

To put Fig 1 into the context of PASS-IT: PASS-IT was initially charged with the delivery of “assessment of learning” which wears a summative tag. But, as students prepared for the various trials much “assessment for learning” occurred as a formative activity for both students and their teachers. Moreover, as the project progressed, the reporting system gave glimpses of students discovering their own learning strategy in line with the “assessment as learning” approach.

References