A unique style of computer-assisted assessment

Mike Thelwall

School of Computing and Information Technology, University of Wolverhampton. Email: cm1993@wlv.ac.uk

This paper examines a project at the University of Wolverhampton that has been producing its own unique style of computerized test for several years. The tests are all designed to deliver a different set of questions each time they are run, a fact which enables many of them to double as learning resources. Most of the tests are used for both formative and summative assessments on Level 1 modules, in conjunction with more traditional assessment methods.

Introduction

The Computer-Based Assessment project at the University of Wolverhampton has produced a number of tests built around a common framework. There are three main tests that are used by about 800 students per year, and they replace written tests. There are also two diagnostic tests that are not part of any formal assessment but are used by about 400 students per year. We have produced numerous special versions of these tests for short courses, different teaching techniques, and one for a Broadnet Online training module (Broadnet, n.d.). The tests are all written with 80,000 permutations built in to allow them to be made available for students to use and practise on at all times without compromising security. In fact, there are no serious security concerns because of the random factors, and so the same tests are used every year without the need to be hidden or rewritten. Similar advantages have been found on other projects using random factors; see for example Thoennessen and Harrison (1996).

The project has been producing random-based computerized tests for modules for several years now. It began with a single statistics test, the success of which led to the use of the technology for the production of others. Our criteria for accepting a module for a computerized test are that it must have a large number of students and that the tests should be able to be used for a number of years. We have produced five random-based PC tests so far, two of which have a number of versions for different modules and situations. Three of these replace written tests in maths, stats and IT, while the other two are diagnostic tests for the numeracy and computing skills of new students.

The main features common to all the tests are as follows.

- Detailed feedback is given at the end of each test.
- Random factors are built in so that each use of the program produces a different set of questions.
- The assessment grade is shown on-screen at the end of the test (for students and instructors) and is saved automatically to a common database together with comprehensive information about the test. Lecturers using a Web browser can access this data to analyse individual student performances or to produce summary statistics.
- Once written, the same test can be used every year without modification.
- The majority of questions are not multiple choice or questions expecting keyword answers.

On a technical level the tests have a number of distinctive attributes, as follows.

- They run in Windows versions 3.1 and higher.
- They are written in a programming language rather than with a specialist quiz authoring tool such as Question Mark Designer.
- Each test consists of one program file and two standard library files, all of which fit on a single floppy disk. There is no data file, the data being stored in the program.
- Data is saved using Internet protocols, a method that works on any computer connected to the Internet, even a student's home PC.

At the start of the project we were tempted to use the Web to run the tests, but rejected this as not yet giving the very high reliability needed for large-scale summative assessments. Others have managed to use the Web for assessing technical subjects in a different way (see Online Exercises, n.d).

Although the same technique and randomization are used for the production of these tests, they fall into three different groups: assessment tests available all the time; assessment tests available only for the actual test; and diagnostic tests which do not count as part of the assessment for any module. The maths and stats tests that are covered in most detail here are available at all times for students to practise on. They count for assessment only when taken in the actual test session. The IT test is password-protected, and the students use it only for the actual assessment test. It was thought that the nature of the module meant that there would be no benefit to students from practising it. This was mainly because they would end up practising the test in the workshop time instead of using the package they were supposed to be learning. The other two tests are diagnostic tests for numeracy and computing which do not count towards assessment. These ask questions depending on the answers to previous questions.

The method of producing the tests has been to obtain a number of past test papers for the module from the appropriate lecturer and to produce a prototype computer test based on these. This is then shown to the module leader for comment and amendments. We found that this method was more practical than involving lecturers in the production of the material at an earlier stage as they often were uncertain about what was possible for their module or how to introduce the randomization.

A comparison of the computerized tests and their written predecessors

The tests produced by the project matched the written tests that they replaced to varying degrees. All assessed the same general learning outcomes as their written counterparts. Some types of question from the original written versions could not be asked in the same way, but many others could. For example, the computerized stats test questions looked very similar to the paper versions but most of the IT test questions changed completely, although testing broadly the same knowledge. There was an attempt to make the computerized questions smaller or to break them up into parts so that there would not be questions or parts of questions with large all-or-none mark allocations. Questions with large mark allocations that could be lost with a single mistake would be unfair to the students.

The exercise of rewriting the tests for the new medium was an interesting one, helping us to rethink assessment criteria and to produce new types of questions. All except one of the tests ended up with at least one question that was some form of multiple choice. Despite this, multiple-choice questions were not extensively used in any test as it took more time to produce a sufficiently large random selection of them than with other types of question.

I shall discuss the particular characteristics of each test in turn.

The statistics test

This test was probably the closest match to the written version, with a similar number and style of questions. Most questions in the original test described a situation, gave some data, then asked for statistical calculations and a report on the results. This was done in the same way in the computerized version except that a report was not asked for; instead a multiple-choice question concerning the conclusions of the report was included on the same screen as the boxes for the answers to the calculations. Some of the aspects of the report that were not tested in this way were tested in additional multiple-choice questions later in the test. Although the learning outcome of being able to write an entire report was not tested, all the individual component parts of the report were asked for in one question or another. Figure 1 shows a sample question from the test.

The Information Technology test

This produced the greatest change from the written to the computerized version, a fact due to the nature of the subject material; I believe that a computerized test is more natural for this situation. The original paper test was based on short written answers that were awkward for computer-marking. We did, however, devise a method of marking short answers with deductive logic. An example of this approach is: 'If they have put the correct answer, then they must have used words A, B and C or D and E, so we will mark it correct if it has A, B and C or D and E in it'. This worked surprisingly well. One of the questions is: 'Describe how to start Word from the Windows Main Menu screen'. There is a box for the answer to be entered, and the marking scheme will give full marks if the answer contains the words: 'double', click' and 'icon', or the words 'click'; 'icon' and 'return', or the words: 'click'; 'icon' and 'enter'. It is hard to enter a correct answer without using one of the three sets of words above. Any students who did so would be caught in the safety net of showing their answer and marks to the instructor at the end of the test to query why they were marked wrong, and to get their marks restored, but this turned out to be a rare

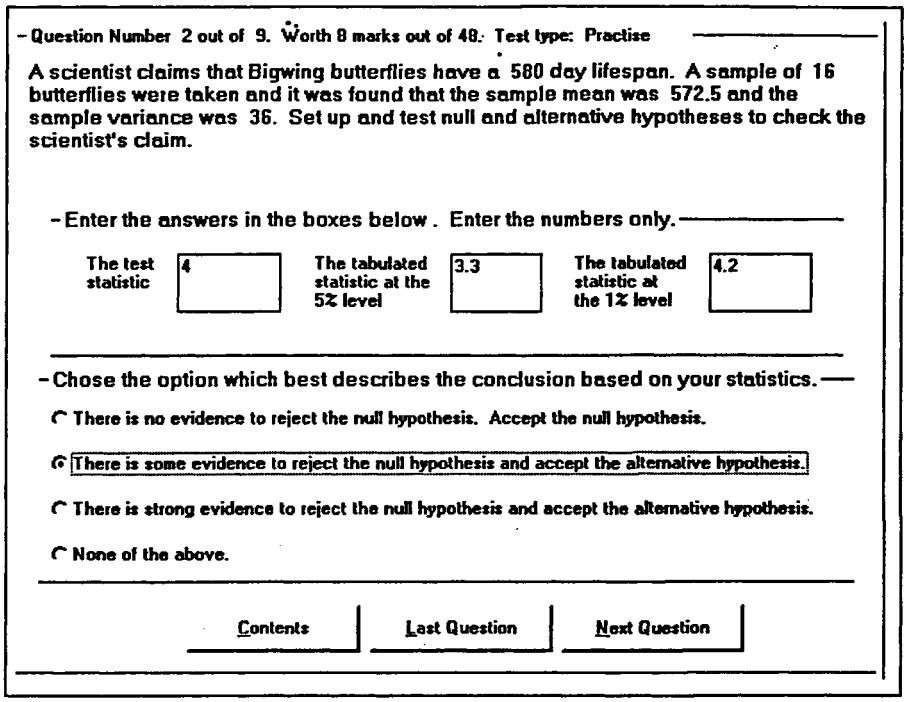

Figure 1: An example of a question from the test

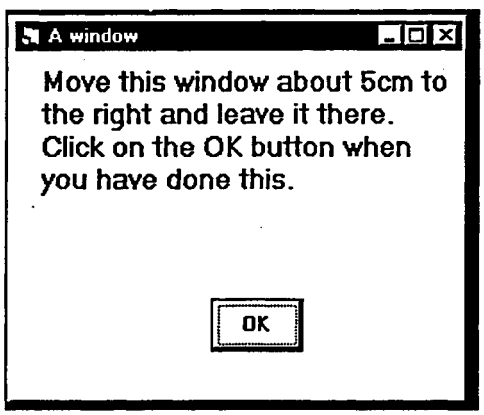

Figure 2: A test of the ability to move a window

event. For this question, the students lose marks if they use any of the words 'drag', 'type' or 'menu' as these could not be used in a correct answer except in very obscure cases.

No serious complaints were made about this system from the students, even after they had seen transcripts of their tests. A check for words which could not possibly be included in a correct answer is made in the short-answer questions to stop students simply writing down as many key words as they can think of. We allocated a third of the marks on the paper to these short-answer questions, and invented new types of question for the remaining two

thirds. These questions were mostly of types that could not be asked in a written test, for example a question testing the ability to move a window with the mouse (see Figure 2)

This summative test counted for half of the mark for the first topic of the module, the other half going on continuous-assessment exercises. This is a combination similar to that used in an IT module at Loughborough, where a computerized multiple-choice test was used in addition to an essay (Stephens, 1994).

The maths test

This test replaced a written calculus and algebra test to which it was broadly similar. The unusual feature of it was the need for extra interfaces to display maths and to accept mathematical inputs from students. This made it a rather more complicated product than the other tests, although it still received a high usability rating from the students.

This question also required a more complicated interface than needed for the other tests. The answer is an algebraic expression that has to be entered with special syntax, then the box at the bottom (see Figure 3) is used to check that the syntax is correct.

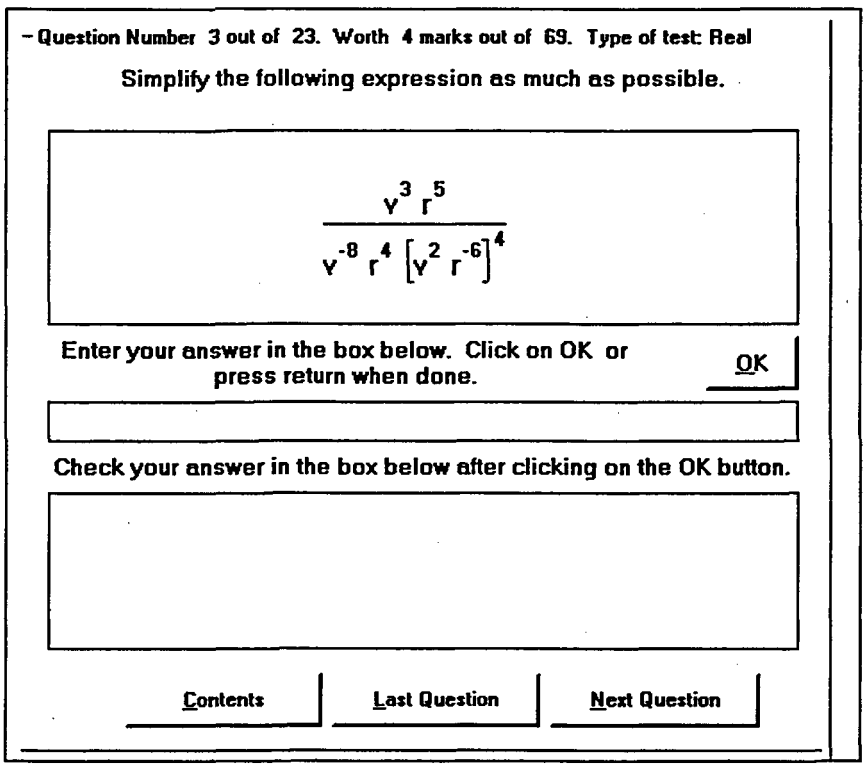

Figure 3: The interface for the maths test

The computing diagnostic test

Computer-aided assessment tests are widely used in computing, for example at Portsmouth (Callear and King, 1997). We produced a computing diagnostic test in response to a request from members of staff who had seen other tests in action. An introductory programming module at Wolverhampton has used a written programming-skill test for several years to allocate students to streamed tutorials. We produced a computerized version of this to save marking time and to allow the test to be extended without creating extra marking. The computer test was again very similar to the written version, but we partly abandoned the questions asking the students to produce a program in any programming language to achieve a given task. Automated program marking is widely used, for example the Ceilidh system (Foubister *et al,* 1997), but devising a system to work for a wide range of programming languages would have been very time-consuming. The programming question was still asked, but was left unmarked, the students being given the option of asking a live tutor to mark it on their transcript if they felt that the mark on the rest of the test was too low.

Questions on the related skill of program comprehension were included, partially to fill the gap. The question transcripts were all saved to enable the tutors to judge whether the absence of marks for these questions was seriously skewing the results of the test. This did not seem to be the case.

The numeracy diagnostic test

This is the only test in our portfolio that did not replace a written test. It was used in three maths modules where it was felt that a diagnostic test was necessary. The need for diagnostic testing had been previously identified but not yet actioned. Maths diagnostic tests have been written by others (Appleby *et al,* 1997; Knowledge, n.d.) with a better theoretical framework, but those we tested had data-saving methods incompatible with our networks.

Analysis of student usage of the maths and statistics tests

I conducted studies on the maths and statistics tests in the first semester of the 1997/98 year, and wrote two internal reports (Thelwall, 1997; 1998). I used three sources of information on which to base the reports. These were the computer logs of the tests, a questionnaire given to the students, and a number of informal interviews. Close analysis of the data is problematical for a number of reasons. One such is that a previous study I carried out showed a high correlation between students' attitudes to the test and their grades on it. This could be due to an appreciation of how to use the test leading to a better mark, or the feel-good factor from doing well on the test. It was probably a mixture of the two.

The following sections summarize the general findings of the reports that are appropriate to this paper.

Student usage of the statistics test

We discovered that the 168 students who took the test were practising it three times on average before taking it for a grade. They found it easy to use and trusted it to mark their work. The students believed that the test helped them to learn and motivated them to do more revision. The overall high grades on the test reflected this belief. The printed

feedback sheet given at the end of the test was particularly popular. Figure 4 shows an example of feedback on one question.

```
4) A biologist was interested in the weights of male and
female Lappin
Rabbits.  She weighed a random sample of both and
calculated the following
sample statistics.
          Sample    Sample    Sample
           Size      Mean    Variance
Male         12       3.1      1.46
Female       11       4.6      1.61

Test a suitable hypothesis.

Your pooled variance was 22.  The correct answer was
1.531.

Your test statistic was 34.  It was wrong.  The correct
answer for YOUR
pooled variance was -.766.  The correct answer was -2.904.

Your tabulated statistic at the 5% level was 4.432.  The
correct answer was
 2.08.

Your tabulated statistic at the 1% level was 5.654.  The
correct answer was
 2.831.

Based on the numbers you entered your conclusion was false.
 0 out of  10
```

Figure 4: An example of feedback on a question

The feedback given is enough for the students to take away and sit the test again at home. Many did just that.

The test was a positive experience for most students, motivating them and helping them to learn. In fact, I believe that it was more successful at teaching than the specially written computerized tutorial they also use.

Student usage of the maths test

The maths test was more technical and more complicated for the students than the stats test. In order to enter some of the answers, students had to learn some syntax for entering mathematical expressions. Despite this, the conclusions from the same data sources as the stats test were broadly similar. The students practised slightly less on average, just under three times, but still felt that they had learned from this practice.

Other tests

The other programs written for the project were not designed to be used for practise for various reasons, although they do contain the same random factors and are built using the same technology. Practising would be inappropriate on the diagnostic tests and formal assessments. As mentioned above, the IT test is password-protected, and this is to prevent practising, but it does count for assessment. No formal studies have been conducted on this test, but I believe that it would get a lower user-satisfaction rating than the other tests partly because it does not allow practising.

Staff attitudes to the tests

One of the main aims of the project was to gain staff acceptance of computer-assisted assessment by getting round some of the problems which have been identified (Bull, 1994). There are two main incentives for lecturers to commission one of our tests: to save marking time, and to have an extra learning resource for the students. Once written, the tests require very little attention. Problems that need to be fixed are sometimes identified, and syllabus changes also cause alterations to the test. If no changes are wanted, the programs stay on the student networks ready for the next run of the module. Staff did not raise any serious concerns about the ability of the programs to assess their students, although a few issues were raised. The short-answer questions were the only problematic ones, the non-transparent marking technique baffling instructors and students on occasion. A more serious problem was that many of the weaker students took too much time on these questions and did not manage to answer all the others. For next year, there will be fewer questions of this type, and they will all be at the end of the test.

Another of the original objectives of the project was to gain long-term use of computerized testing. None of the tests has yet been dropped from a module, a fact that I see as going some way towards the achievement of this objective. I attribute this staying power to successes with students as well as ease of use for staff. In fact, it is true that dropping a computerized test will result in an increase in work, both from the need to write a new test and from having to mark it by hand.

Conclusions

The project has produced programs that have stood the test of time, and is continuing to produce more with the same methodology. So far, each test has been easier and quicker to produce than the previous one. The random base allows the tests to have long lifetimes and avoid security problems in most cases. It also allows them to be successful as learning resources and thus helps them to be popular with both staff and students.

I do not believe that there is a significant difference between the ability of the written and computerized versions of tests to assess skills and knowledge. A computerized test, seen in isolation from the way that it is used, is in general a somewhat blunter instrument. It is unable to cope easily with such things as bad spelling, less able to give 'method marks', and is restricted in the type of questions it can ask. These problems have been circumvented to a large extent by the use of appropriate question designs. The tests also contain additional features in the feedback generated and in the ability for students to have marked practices that are hard to achieve with written tests.

The tests are available for others to use without charge from the Web site (http://cba.scit.wlv.ac.uk/).

References

Appleby J., Samuels P. and Treasure-Jones, T. (1997), 'A knowledge-based diagnostic test of basic mathematical skills', *Computers in Education*, 28 (2) 113–31.

Broadnet (n.d.), *The Broadnet Project*, http://www.broadnet.co.uk.

Bull, J. (1994), 'Computer-based assessment: some issues for consideration', *Active Learning* 1, 18–21.

Callear, D. and King, T. (1997), 'Using computer-based tests for information science', *ALT-J*, 5 (1), 27–32.

Foubister, S.P., Michaelson, G.J. and Tomes, N. (1997), 'Automatic assessment of elementary Standard ML programs using Ceilidh', *Journal of Computer-Assisted Learning*, 13 (2), 99–108.

Knowledge (n.d.), *The Knowledge Space Project and ALEKS*, http://aleks.uci.edu/ and http://www.spaces.uci.edu/.

Online Exercises (n.d.), http://math.uc.edu/WWW-test/demo/demo.html.

Stephens, D. (1994), 'Using computer-assisted assessment: time saver or sophisticated distraction?', *Active Learning* 1, 11–15.

Thelwall, M. (1997), *A Study of Student Use of the Statistics Computerised Assessment Test*, internal report, Wolverhampton University.

Thelwall, M. (1998), *A Study of Student Use of the Mathematics Computerised Assessment Test*, internal report, Wolverhampton University.

Thoennessen, M. and Harrison, M.J. (1996), 'Computer-assisted assignments in a large physics class', *Computers in Education*, 27 (2).