ASSESSING THE USE OF A NEW QTI ASSESSMENT TOOL WITHIN PHYSICS

Richard Bacon

Assessing the Use of a New QTI Assessment Tool within Physics.

R A Bacon
Department of Physics,
University of Surrey,
Guildford, Surrey.
GU2 7XH.
r.bacon@surrey.ac.uk

Keywords interoperability, science, number

Abstract

Computer assisted assessment has been used for several years in the Physics Department at the University of Surrey for routine coursework assignments in some first year modules. The tests have been administered using the SToMP (Software Teaching of Modular Physics) testing system that was developed in 1994 as an integral component of the SToMP Multimedia learning and teaching package. A completely new SToMP testing system has recently been created that is designed to be IMS-QTI compliant, and it is able to offer some new features as a result of its interpretation of the QTI standard. The application of, and student responses to, some of these features will be discussed in this paper as well as an important extension to the standard that is being proposed to improve the setting and handling of numeric questions.

Question types that are available in both SToMP systems and that have been used in the tests at Surrey, include numeric questions with randomised values as well as the more conventional types. A feature of these numeric questions is the ability to assess both the accuracy and the precision of the students' responses. The new testing system was introduced last Autumn and one of its innovative features was to allow the students to return to each test after it had been marked for the whole class, and see both the marks they obtained and some formative feedback.

A questionnaire was designed to assess how the students viewed this system, and how they felt about having coursework dealt with in this way. Some of the responses from this questionnaire display strong antipathy to the system and a serious lack of faith in the marks obtained. Some of the results from this survey will be presented and discussed, particularly with regard to how methods of setting of such coursework using CAA could be improved to address the issues raised. The discussion will include a resume of how such changes might impinge on the new interoperability standards.

Introduction

The testing system used in the work described in this paper was created over the summer of 2002 and was used with students in the autumn of 2002. It was created to replace an aging testing system used in the SToMP package (Hunt, 1997) and it was designed to be IMS-QTI compliant, working directly from the xml. It is not web based, but runs native on a PC obtaining questions from, and writing results to, a database over the internet.

The system was designed to be IMS-QTI compliant in order to facilitate the exchange of questions, and to improve the marketability of the system. Most of the realisable combinations of the five QTI response types (logical identifier, numeric, text, x-y and group) and the four input mechanisms (choice-list, text-box, slider and hotspot) have been implemented. These support most of the types of question that are needed in the sciences and that are available in such a standard. Further question types are being supported as the implementations of other combinations become clarified (Barr 2003) and as extensions to the standard are developed.

Support for numeric questions

One of the most significant shortcomings of the current QTI standard is in the handling of numbers. Numbers entered by student users can be tested for equality and for being larger or smaller than absolute literal values, but the format, precision and relative accuracy of such input cannot be tested. Neither is there any provision within the standard for the randomisation of numeric values used in questions. It was felt that these shortcomings of the standard would have a damaging effect upon its use in the sciences, and so an extension has been drawn up (CETIS 2003), implemented, tested and submitted to IMS for consideration in the next revision.

The extension, originally called 'random numbers' but now renamed 'question variables', involves three new xml elements. The first, <questvar_create>, allows a variable to be created and given an absolute value, a value randomly chosen from a range of values, a value randomly chosen from a list of values, or the value of an arithmetic expression that can involve other variables. Other information can be included, such as the number of significant figures, decimal places, whether it is to be represented in e-format or in a different number base. The second element, <matquestvar>, formats the variable as text for output to the screen as part of a question or a feedback message. The third element, <questvar_equal>, allows for the testing of user input against one of the question variables according to the precision of the user input (e.g. the number of significant figures) and the accuracy of the value (either relative accuracy or absolute). This extension has been implemented in both the SToMP system and the CETIS standard rendering tool (CETIS 2002) and the former has been successfully used with students.

Using question variables.

The first major use of this system was with a first year undergraduate Physics course in Data Handling. Nine tests were set, each of about seven questions. Five different styles of question were used, single and multiple choice from a list, pair matching (pairing items from two lists), sorting (putting items into a rank order) and numeric. These tests formed the main assessment of the course. They were spread out over nine weeks, and could be done when the students liked (within the time constraints of the test scheduling). Due to the fireman's strike coinciding with these tests the availability of our departmental computing lab was seriously restricted. A downloadable version of the testing system was therefore made available to the students so that they could take the tests on their own computers at home, or in other computer labs within the university.

Under these circumstances, the randomisation of questions and numbers used in the tests was felt to be of utmost importance. The testing system allowed alternative versions of a question to be created from which one could be randomly selected for each student (not a QTI feature), and this randomisation was used for most of the single choice and multiple choice questions. Numeric questions, however, benefited from using the number system described above, so that each student saw the same question but with different numeric values. This feature was also used with some of the other question types. Where alternative values were presented in a list of options, for example, these values were usually randomised as well as the order of the list itself being randomised (as supported by QTI).

A questionnaire was designed to evaluate the students' opinions about this new testing system and it was distributed to all the students. It proved very difficult to get the majority of students to actually complete and return these. Eventually 38 responses were obtained from the 55 students who took the tests. From these responses it was quite clear that some of the students did not realise that these tests formed the main assessment for the course, whereas others seemed to confused them with another electronic test they had taken using the old SToMP system. The overall trend of the responses was clear, however.

The questionnaire contained both semantic differential style questions and free text entry boxes, with some questions inviting both types of response, and some just asking for text. The results reported here are a distillation of all the free text entered on each questionnaire, according to the perceived meaning of the points being made.

None of the following points were suggested to the students within the questionnaire.

	Number of students	point being made
a	23	Did not like being unable to show working and getting marks for it if their answer was wrong.
b	20	Liked the flexibility in timing of the taking of each test.
c	12	Liked the automatic marking.
d	10	Did not like the 'all or nothing' marking.
e	8	Found the system easy to use.
f	6	Found it less stressful than conventional tests or exams.
g	5	Did not like having to take each test in one go.
h	5	Did not like the test system appearing to 'go wrong' at times.

Overall, there were 56 points made in favour of the system and of using it, and 78 points made against the system and its use.

The testing system had some features that need explaining in order to understand the significance of some of the comments.

The first such feature, which is believed to be innovative, is that the system can provide formative feedback to the student after the 'end date' of a summative test. To obtain this feedback a student must start the test again during the feedback period. The original questions are displayed together with the responses the student made, but in a read-only form. When the mark button is pressed for each question they see the available formative feedback for their answer to that question as well as the mark that was awarded. Students can only see their own marks. The last point (h) in the table, that the testing system appeared to 'go wrong' at times, referred to the information given in the feedback system. One of the questions was originally authored with the wrong marking algorithm and thus gave the wrong answer and the wrong marks during the feedback period. This was reported and corrected, but some students either were not aware of this correction, or were sufficiently upset by it happening at all that they mentioned it on the questionnaire. There was only one other fault that the author was aware of, and this was reported by only one student, was corrected the same day it was reported, and did not affect the student's marks.

Point (g) is also interesting as it is not strictly a feature of the system. A test can indeed be interrupted and resumed, so that if a machine crashes or the network connection is broken, a student is not disadvantaged. The students were not told of this since we did not want them deliberately crashing their machines. Following this comment in the questionnaire responses the ability to interrupt a test will be made a full feature of the system that will be available if the course tutor feels it appropriate. In the case of these assessed tests, however, the tutor has already indicated that he would not want them to be interruptable in this way.

Points (b), (c), (e) and (f) were encouraging. It was not clear what feature of the automatic marking was being appreciated, objectivity or promptness. Only one student reported that they found the tests stressful, although there was evidently some confusion over what the CAA was being compared to, i.e. conventional coursework or an examination.

Addressing the issues.

The two most frequently mentioned negative points (a) and (d) involved the way that numeric questions are marked. The all or nothing marking and the inability to cope with working are clearly linked, although the former could also refer to the way that the marks were actually awarded. This needs some explanation to put these comments into context.

Simple numeric questions were awarded one of four different marks. Typically a correct answer (correct value and the correct precision as specified in the question) would be awarded 5 marks. If the answer was only approximately correct (usually within five percent above to five percent beneath the correct value) the student would lose one, two or three marks. If the precision with which the number was specified was wrong (but the value was correct) then again they might lose one, two or three marks. The number of marks lost in each case was set by the course tutor. If the answer was wrong (i.e. it did not match any of the above criteria) then it was awarded zero marks.

Such questions might involve a simple application of a formula, typically involving three or four values given in the question or implicit to the formula being tested. The method taught in such cases is to defer calculation until all the numeric substitutions have been made, in order to minimise numerical error. In this sort of question the only working that could yield more marks for a wrong answer would then be evidence of the correct application of the formula.

A solution might be to invite the student to enter any working in a text input box which could be assessed by the tutor if the numeric value was wrong. The support of on-screen editing of algebraic expressions is not straightforward, however, and much of the point of the automated system would be lost with such a scheme, and so this has not been followed up.

In five of the tests there were questions which involved more complex processes, such as 1) taking the sum of a set of twenty numbers, 2) finding the mean and 3) squaring it, then 4) finding the sum of the squares of the twenty numbers and 5) finding its mean, and then 6) subtracting one from the other and 7) taking its square root. Questions like this were split up, and required the student to enter the value of each section, as illustrated above. Whilst the working was not displayed, each section involved only a relatively small calculation. It was true, however, that if a student made an error in the first part, and propagated this error through subsequent sections (4 for the sum or 3 for the sum of squares) then they would lose marks for each affected section.

One test involved a question that had to be answered using excel. This required several different analyses to be applied, and several different questions answered from the same set of data.

The interpretation of the students' comments in the light of the actual questions which they were asked is therefore not as simple as might initially be thought, particularly since a number of students showed confusion over which test was being surveyed. It would seem reasonable, however, to treat the criticisms as generic and to try to address them however they occur.
In response to the criticism, two changes have been introduced into the system based upon further extensions to the QTI standard. One of these is quite simple and has already been used in a second year test but the other is rather more complex.

Including student responses

The type of multi-part question, described above, that requires students to enter numeric answers to each part of a compound problem can be improved by allowing an incorrect response to one part to be used as a basis for the answer to subsequent parts. Such a system is not conceptually new (Ashton 2002), but it is not supported within the QTI standard or in any system the author has seen.

In order to support this functionality within the proposed question variable system described above, all that is needed is one additional xml element <questvar_update> that can be included in the response condition testing. This element allows a previously defined question variable to be initialised to a value entered by the user if certain answer conditions are met (e.g. if an answer is wrong, or inexact), or to be initialised to the value of an arithmetic expression involving other variables (as before).

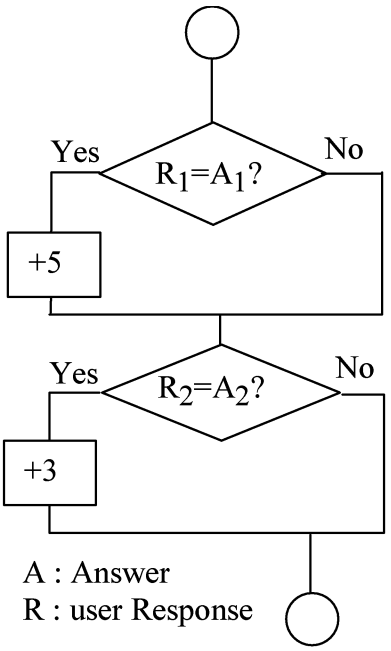

Figure 1

Flow chart schematic of the conventional marking of a two part question.

Incorrect, or partly correct, answers can then be used as the basis for calculating subsequent answers in multi-part questions. Consider a question comprising two parts, the first part asking the student to calculate the average of a number of items, and the second part to calculate the square of that average. The first part is to score 5 points, the second part 3 points. If the student enters the wrong value for the first part, but correctly squares that value and enters it for the second part, then it would seem reasonable to award zero for the first part and 3 marks for the second part. Figures 1 and 2 show flow charts of how the

answer processing might look without this facility (Figure 1) and with it (Figure 2).

In Figure 1 the student response to the first part of the question (R_1) is compared to the correct answer (A_1). If they are equal (i.e. the answer is correct) then 5 marks are added, if they are not equal then no marks are added. The marking then proceeds to the second part, and the student response (R_2) is checked against the correct answer (A_2 {= A_1 * A_1}) and 3 marks added if they agree.

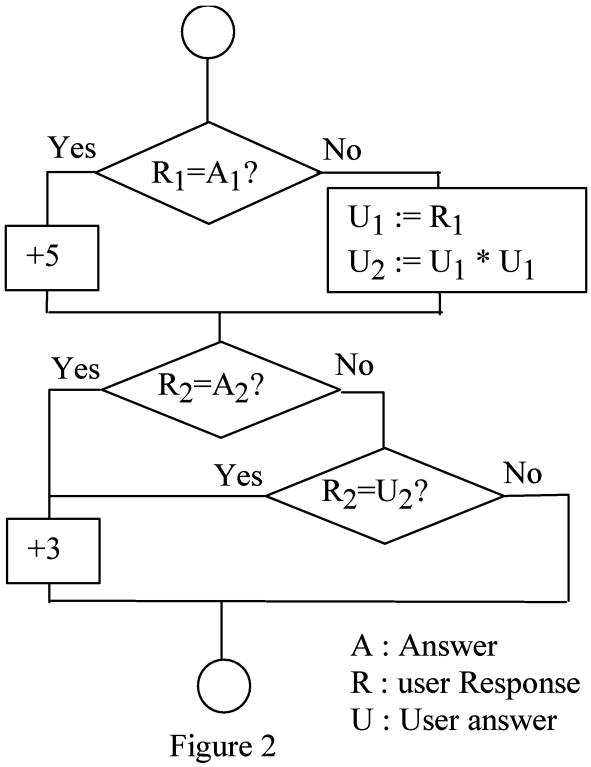

Figure 2

Flow chart schematic of the marking of a two part question, where the user's answer to the first part can be used to create an alternative answer to the second part.

Figure 2 shows the additional processing necessary for checking the students response to the second part against his own response to the first part. If and only if the first response does not match the correct answer, the user's response is assigned to a new variable U_1. A second variable U_2 is then assigned the value of the square of U_1. If the student's response is equal to A_2 then 3 marks are awarded as before, but if the response is not equal to A_2 then it is checked against U_2 and if found equal the 3 marks are awarded. The variable U_1 can be used later, in formative feedback, to clarify the criteria against which the response to part 2 is being checked.

This example and the flow charts are considerably simplified, both in the nature of the operation of the testing system and in the checking that would be applied to the users response. They are sufficient, however, to illustrate the principles.

This system has been implemented, and will be used in the data handling tests described above, for the multi-part questions.

There are two obvious disadvantages to this scheme. The first is that the student is not having to make decisions about how to go about solving the whole problem, because it has already been split into manageable chunks by the question author, the second is that in some problems the accuracy of the final result will be compromised by the rounding errors introduced as the calculation proceeds in small increments. Students are taught to leave all

calculations until the complete expression for the solution is developed, if possible. The expression can then often be simplified (reducing complexity and calculation error) and, even if this is not so, maximum precision can be maintained throughout a calculation more easily if it is carried out as a single task. This approach is clearly not encouraged by splitting a problem up as described above.

Tutoring through a problem

A second approach to assessing a student's ability to solve complex problems is to offer the student alternatives at each step of the problem, to provide feedback as to how appropriate the most recently made choice is, and to record each step taken.

Consider the problem of adding two 2D vectors. This can be done in a number of ways such as by resolving the vectors into components in two orthogonal directions or by forming a vector triangle (or parallelogram) and calculating the resultant using geometry. The options for the first step following the question statement could be, for example, to resolve into components, to take moments (as a distracter), or to form a vector triangle. If the student chooses to resolve the vectors into components, then the next step could be to choose the two directions for the components - the most suitable may well not be horizontal and vertical. If taking moments was selected, then an explanation of why this is not the best method could be given.

Whilst this still decomposes the problem into many small steps, it does it in manner that leaves the student with many choices to make, and also does not necessarily involve partial calculations. Steps can offer lists of expressions rather than asking for numeric values, and the calculation itself might be redundant, or could comprise the final step.

This sort of tool is not new, and a machine implementation in this manner is not far from the sort of machines discussed by Skinner in the 1950s. A tool already exists within the SToMP teaching and learning system that does precisely what is described above. It is a stand alone tool, however, that does not record any of the steps taken by the student and it is very time consuming to set up even fairly simple problems because of the very large number of possible steps that need to be addressed. A problem in which the resultant of three forces was to be found needed to 66 different response pages, even though several groups of 'bad' choices were dealt with by generic pages. For want of a better name, and to conform with existing SToMP nomenclature, the system described here will be called a 'problem tutor' in the remainder of this paper.

The purpose of describing such a tool here is to show that it is a relatively simple matter to extend the QTI standard to duplicate this functionality, largely because of the extremely flexible and comprehensive way in which the standard has been developed.

The QTI standard supports the ideas of sections and items, and one section can contain several items. It is not unreasonable to equate an item with one screen view - which might be one question, several related questions, or one question with several parts. If each step in the problem tutor is dealt with in one item and all the items concerned with one problem are contained in one section, then if the next item to be displayed at each stage could be made a function of the user's response to the current item (rather than a simple sequential progression as at present) then a problem tutor could be synthesised from a QTI compliant testing system.

This can be done by defining one new xml element that goes within the conditional tests (of the users responses) and defines the next xml item to be displayed if the condition is true. The SToMP QTI testing system has already been modified to allow for this new element, and some simple examples have been prepared. The flexibility of this development as both a learning tool and as a discriminating assessment tool is exciting and, because it is evolving from a standard testing system, with its alternative styles of user response, the potential is far greater than the earlier SToMP problem tutor system. It is anticipated that with well designed questions, students will be better able to gain credit for their problem solving abilities, rather than just for being able to produce accurate numerical answers.

Conclusions

It is clear that the proposed question variable extension to the QTI standard only addresses some of the problems of the use of CAA in the numerate sciences. Further extensions will be needed in order to more closely match student expectations for marking in the sciences. Implementations of two such extensions have been described here. In each case the extension is relatively minor, but greatly extends the functionality of the system. Considerable additional work will be required, however, before questions based upon the use of such QTI compatible tools will be able to compete seriously with conventional paper based examination and test questions in subjects such as physics.

Acknowledgements

The author would like to acknowledge the contribution of Dr Graham Smith (Leeds) to this work, and for his alternative implementation of the question variable proposal.

References

Hunt, J.L. and Bacon, R.A. (1997) *SToMP: A dynamic Approach to Courseware Development*. Proceedings, ED-MEDIA 97, Calgary. 670-675.

CETIS (2003) *Question variables (Dick Bacon, University of Surrey)* <http://ford.ces.strath.ac.uk/QTI/working_papers.html> (May 2003)

CETIS (2002) *CETIS QTI rendering tool (Graham Smith, University of Leeds)* http://www.odltest.leeds.ac.uk/QTHTM/SIGDEMO/sigintro.htm (Sept 2002)

Ashton, H.S. and Beevers, C.E. (2002) *Extending flexibility in an existing on-line assessment system* , Proceedings 6th International CAA Conference, Loughborough University.