DISCRIMINATIVE MARKING OF NUMERIC PROBLEMS

Richard A. Bacon

Discriminative Marking of Numeric Problems

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

Abstract

In previous papers (Bacon 2003, 2004) the author has described some of the results obtained from surveys of the use of the SToMP testing system for the coursework assessment of one first year and one second year module within a Physics Degree programme. This paper will deal with progressive modifications that have been made as a result of student feedback from these trials and with preliminary analysis of the feedback obtained from the students using the updated tests.

The SToMP testing system was written in 2001/2 as a direct implementation of the IMS-QTI v1.2 specification, but includes several extensions for handling numeric problems of the type frequently found within science and engineering courses. Such problems typically require a numeric answer to be judged by its precision (e.g. the number of significant figures) as well as its accuracy (i.e. whether the value falls within a specified range). The system must also be able to recognised alternative forms of the same value and precision in scientific format. These features were mapped onto a suite of extensions to the QTI specification for ease of implementation, and included other features such as alternative number bases and the randomisation of values within questions.

One of the features supports the propagation of errors in multi-part numeric problems. A wrong answer to an early part of the problem is remembered by the system and used to generate alternative answers to later parts. This means that the student making such a mistake need only be made to lose marks for the part in which they made the mistake, not in later parts. A previous year's trial of this feature, although it contributed to student's marks, was not appreciated by the students because they had not been informed of it. The results of this year's trial will be reported, where students were informed (despite academic misgivings that this would affect the care with which they prepared their answers).

The type of numeric problem mainly dealt with in this paper, is where one or more formulae have to be identified by the student as being appropriate, they have to be solved for the parameter required by the question, suitable values have to be substituted for the expression's parameters and then a final value calculated. A previous paper (Bacon 2004) described a partially successful strategy for assessing the quality of a student's working when they arrived at a wrong final answer. This involved the student entering the numeric expression from which they obtained the answer, as well as the final value itself. Half the respondents to the survey of this system complained of "no marks being available for their working", which was not unreasonable since the expression was only evaluated. This paper will describe further work that has been done on the analysis of the structure of the students' expressions. This has led to more detailed marking and more helpful feedback being available to those students who do not get their answer or their expression correct. The system will be described and students' perceptions will be reported.

Introduction

The SToMP QTI assessment system has been used with students since 2002, as described in earlier papers (Bacon 2003, 2004). The system is IMS-QTI V1.2.1 compliant (IMS 2003), with extensions to support question cloning and the scientific use of numbers (Bacon and Smith 2003).

It has been used for coursework assessment in three first year physics undergraduate courses, one second year course and one MSc course, although the work reported here involves just one of the first year courses (data handling) and the second year course (radiation detectors). The overall aim of this work remains the re-implementation of paper based coursework in an electronic form that is acceptable to students, and each year there has been an iteration of the cycle of use and survey of use. Modifications are introduced as a result of the previous year's survey responses, analysis of the student's responses in the tests themselves and study of the student's learning requirements.

The original reason for the introduction of electronic methods for coursework was to reduce the marking workload of academics and to provide the prompt feedback of marks and comments to students whilst keeping the student's workload about the same. The reason for the coursework is to provide students with exercises relevant to the topics being covered in a lecture course, and the reason the coursework is marked is largely to ensure that students actually do the work. It is beginning to emerge that electronic coursework of this style may be able to offer more than the paper based equivalent

Survey results

The table below gives a summary of the comments made in the students' responses to the surveys carried out after each use of the system. None of the questions except in the final 2RD survey actually mention specific features of the system, and so the frequencies with which a feature is mentioned is a fair indication of how much feeling there is about that feature. The table lists all comments mentioned by more than 10% of the students in any of the surveys.

It is clear from the figures that the major objection to the use of the electronic system was its inability to mark the working of a problem. An attempt was made to address this issue for the 2RD course in 2003. As previously described (Bacon 2004), this scheme invited students to enter a numeric expression for their solution in addition to their final value. If the final value was wrong but the expression evaluated correctly, then they were awarded half marks. Although the scheme was explained to the students few saw it as addressing the issue of giving marks for working.

	2002	2003	2003	2004	2004
	1DH	1DH	2RD	1DH	2RD
numbers of responses/students	39/53	29/53	18/37	21/34	22/32
Negative comments					
no marks for working	60%	55%	89%	58%	19%
all or nothing marking	26%	0	0	0	0
don't like doing tests this way	0	0	22%	0	12%
cannot interrupt test	13%	17	n/a	8%	12%
question errors	13%	0	0	0	0
technical problems	10%	31%	0	19%	8%
feedback too general	0	10%	0	15%	0
it encourages cheating	3%	10%	6%	12%	0
have to use laboratory	0	7%	17%	15%	0
problems entering expressions	n/a	n/a	6	n/a	12%
Positive comments					
flexibility of timing	51%	59%	44%	46%	35%
test can be interrupted and restarted	n/a	n/a	33%	n/a	n/a
speed of marking/feedback	30%	0	0	19%	0
prompt feedback and re-try	n/a	n/a	n/a	n/a	19%
good aid to learning	0	0	22%	0	31%
easy to use	21%	17%	11%	0	8%
less stressful	15%	7%	0	8%	0
open book	10%	0	28%	0	23%
entering expressions useful	n/a	n/a	0	n/a	16%
good quality feedback	0	7%	0	15%	0
does not take long	0	3%	0	12%	0
saves academic staff time	0	0	17%	19%	0
getting marks for method	n/a	n/a	0	n/a	4%

Table 1. Frequencies with which each topic was mentioned in free text responses in surveys after each use of the system. 1DH = Level 1 Data Handling. 2RD = Level 2 Radiation Detectors. Years refer to academic years. e.g. 2003 = 2003/2004, etc.

There are some anomalies in these figures, some of which stem from an apparent lack of appreciation by the students that these 'tests' are just part of their coursework and do not replace paper based tests. Thus, the flexibility of the timing of the computer assessment is not really a genuine factor. In a similar way, some students are seriously concerned about what they consider to be the 'cheating' of some students who work together on these assessments. From an academic point of view, that students should work together on an item of coursework is not a problem, so long as they each finally answer the question in their own way. In these computer based tests they are assisted in this by there being randomised factors in almost all questions. Simple item selection questions usually come with three variants, chosen at random as the test is presented. Also, items in lists are randomly ordered so that any discussion must be about the actual item rather than its position in the list. Some of the selection type of questions (single, multiple and pairing) have numbers in the lists and these are usually randomised, and numeric questions are almost all randomised.

There is a growing feeling that the word "test" that is built into the testing system in a number of ways (e.g. the desktop icon to start the system is labelled "start tests") should be changed to something less aggressive, but it is not certain that any one word will be appropriate in all cases.

The testing system is capable of being stopped mid-test and being restarted without loss of data. Whether this is to be allowed in any particular case is an option set by the tutor and for the first year tests it was decided not to allow this. This is reflected in the comments. The second year test was set as interruptible in 2003 and the feature was appreciated. Unfortunately, for technical reasons it was not possible to allow the same test to be interruptible in 2004, due to the introduction of the multiple-try feature described below. It is hoped that this problem will be fixed before use in 2005.

The large number of systems problems that the students have complained about have mostly been due to technical problems in our undergraduate computer laboratories and the networking, although a few have been due to incorrectly prepared tests.

The conclusions drawn from the surveys of 2002 and 2003 were that

a) more explanation should be given about how marks were awarded and what features of an answer can be assessed by the system,
b) better diagnostic error messages should be given for expressions that are entered with syntax errors.
c) some sort of analysis should be carried out on the entered expressions, so that marks could be given for partly correct expressions.

One further change that was decided upon following study of the types of wrong answers that students submitted, was to add a test for a value to be in error by one or more orders of magnitude (e.g. ×10, ×100, ×1000 etc.).

Changes

The first of the changes mentioned above involved just the wording of the questions and was easily carried out. Most of the information was put on the header question-page that preceded the questions proper in each of the tests.

Information about the syntax of expressions was put on both the header page and at the bottom of each appropriate question.

The syntax checking of expressions required more extensive changes, and the expression handling software has been largely re-written. The number and specificity of the diagnostic messages has now been considerably increased.

The nature of the analysis of the entered expressions was decided upon following a careful study of the expressions entered by students in the 2003 tests. Each incorrect expression was hand marked, and the features that obtained marks were noted. A method of processing the expressions was then devised that allowed these features to be recognised.

The expressions are requested from the student in a numeric form, with the values from the question being entered exactly as displayed. The expression processing starts by recognising these values and replacing each of them with an identifier. All other values (except powers) are then reduced to unity and the expression simplified. This process is somewhat analogous to the "stemming" of words (Porter 1980), and allows the relationships of the question values within the expression to be checked, irrespective of incorrect constant values or unit scaling factors. Thus, for example, it would be possible to selectively recognise A+B/C or B+A/C, even if B and C were values given in the question and A a physical constant that had been quoted incorrectly.

Multiple tries

One of the reasons for the original use of this system was to do with the prompt feedback of marks and comments to students. The system has special features to support this and they were described in the first of this series of papers (Bacon 2003). Following the period during which students could take the test there was a period when they could re-start the test and it would give them their marks, together with any comments provided by the tutor relevant to the errors they had made. Whilst this did indeed provide more rapid feedback than was generally the case for hand marked assignments, it was still slow in terms of the students' thought processes. One of the criteria for good learning laid down by Chickering and Gamson (1987) is the promptness of the feedback, and in the case of these numeric problems it was felt that only immediate feedback (i.e. within seconds of an answer being submitted) would be really effective in aiding learning.

For other question types like multiple choice, it was felt that immediate feedback would not be popular with academics since it would probably lead to serious cheating. In numeric questions with randomised numbers, however, it is a different matter. Each student still has to work out his or her own answer value from his or her own question values. Whilst considering how this might be implemented it was conjectured that students could be encouraged to re-engage with a question and apply the feedback if they were immediately offered another try. The success of such a scheme would, of course, depend upon the specificity of the feedback and whether it could be worded so that a student could recognise the error they had perpetrated without being told explicitly how to solve the problem. This is not a new idea, the AIM project (Strickland 2002) has allowed multiple tries for some time, but it is thought to be new within the context of a system supporting such discriminatory marking and feedback for numerical problems.

The marks available to the student for any one question obviously need to be reduced at each try, and if their current mark cannot be bettered then a further try is not offered. In practice this should mean that their error was only in either the accuracy or the precision of the value they entered, and this would mean that they already understood the correct method for solving the problem. A reduction of 25% of the marks available at each try was found to match these criteria for the marking scheme used.

Implementation

The ability to offer students multiple tries at a question was handled by having several copies of the same question, but "hiding" all but one of them at any time. This allowed a simple approach to providing different feedback to the same error in different tries, allowed the marks to be different for the different tries, and also helped keep track of what the student had done, and what expressions and values they had entered. Two new features were required as extensions to the QTI specification

a) the ability to declare an item "hidden" (i.e. not to be displayed when browsing through the questions) until "revealed" during the appropriate result processing, and
b) the ability to re-use a response variable in a subsequent item, and to assign a new value to it only if the new value is larger than its current value.

The "hidden" feature was achieved by a new attribute of the "item" element, and a new element used within a "respcondition" element. The response variable re-use was achieved by a new attribute in the "decvar" element and a variant of the "set" element that would not allow the value of the variable to decrease.

Using this scheme, the reporting from the test contains the mark a student obtained for each try at a question, as well as the expression and the value they entered for each try. The randomised values (which remain the same for each try) are also recorded. It would be possible to use a different set of randomised values at each try, but it was considered to be less likely to maintain the engagement of the weaker students (who stand to gain most from this scheme) if the whole calculation had to be started again at each attempt.

As an example of the sort of errors that can now be detected, the following list is for one of the questions:

- the value is approximate or imprecise

- two specific errors recognised from specific wrong values
- the value is exactly or approximately in error by one or more orders of magnitude
- the value is wrong, but the expression gives a correct or an approximately correct value
- the expression is of the right form, but its value suggests the wrong units have been used for a particular parameter
- the expression is of a form suitable for calculating a different parameter to that requested
- the value is wrong and the expression is wrong by orders of magnitude
- the expression uses the given values correctly, but it produces a wrong value
- the form of the expression implies a particular term is missing

There are 17 different error message provided for this question using combinations of these criteria, all of them ending with a suggestion as to how the student should proceed in order to find out how to do the problem properly.

Results of tests

The features described above were used in the 2RD coursework test used in February 2005, taken by 32 students. The marks obtained for the questions offering multiple tries are given in table 2, below.

question 2			question 5					question 8				
freq	try 1	try 2	freq	try 1	try 2	try 3	try 4	freq	try 1	try 2	try 3	try 4
	12	**9**		**20**	**15**	**12**	**9**		**40**	**30**	**22**	**16**
1×	6	9	4×	12	15			1×	24	30		
3×	3	9	4×	0	15			1×	20	30		
2×	0	9	1×	9	9	12		9×	0	30		
			1×	0	7	12		2×	0	24		
			1×	12	10	12		2×	0	0	22	
			1×	12	9	7	9	1×	0	21	0	16
			1×	2	0	4	2	1×	0	0	15	13
								1×	0	0	15	0
								2×	0	0	0	0
								1×	0	0	0	

Table 2. Marks obtained for multiple tries at questions 2, 5 and 8 in the 2RD test. The figures in bold are the maximum marks available for each try. The final mark awarded in each case is underlined.

It is noteworthy that there was only one instance of a student failing to take advantage of an offered try. Clearly there is a strong compulsion to continue, which probably means that students are willing to remain engaged with a problem if they receive constructive help and a chance of improving their score.

Question 2 is the simplest of the three questions, and everyone who attempted it ended up with the correct method.

Question 5 was more difficult. It is clear that most of those who failed to get the question correct at their first attempt have persevered and, either using the feedback provided or by other means, have improved their attempts. Those getting 12 marks at their first attempt were told (according the value they entered) either that they had probably used the wrong units for a fundamental constant, or were trying to find the wrong parameter. This was sufficient for four of them to come back with the correct calculation at their next try, and the other two students obtained the correct solution eventually. Only one student failed to learn how to answer this question.

Question 8 was the most difficult, but 14 students were able to progress to the correct solution. In the previous year, all but one student had obtained zero for this question. The lecturer of this course was aware of the difficulty and dealt more thoroughly with the topic in 2004, but there were still 15 students who would have failed to achieve any marks without the multiple tries feature, and who would thus have learnt little from the experience.

Student's responses

As can be seen from table 1, 19 students said that they liked the prompt feedback and the ability to try the question again. This ties in quite nicely with the number of student who were able to gain from the feature. Just one student stated that he found the feedback to be unhelpful.

The author had been concerned that the whole idea of students being able to re-try questions and thereby improve their scores would be considered unfair by other students who did not need such facilities to obtain good marks. There is no evidence in the survey responses, however, that any student felt this way.

Consider the figures for the "no marks for working" row in table 1. Let the cohort of students taking 1DH in 2002 and 2RD in 2003 be called "CA", that taking 1DH in 2003 and 2RD in 2004 be called "CB" and the final cohort be called "CC". Now, 60% of CA mentioned the lack of "marks for working" when taking 1DH, and 89% of them mentioned it when taking 2RD - an increase of about one half. On the other hand, while 55% of CB mentioned it for 1DH, only 19% mentioned it at 2RD - a decrease of about one half. The only relevant differences between the tests taken by these groups were the extra information about how the marks were awarded and the improved feedback and multiple tries for 2RD taken by CB. Note that CC has not shown a marked change in the proportion mentioning the lack of marks for working. It is therefore considered likely that the drop in the proportion of CB mentioning "marks for working" is due to the improved feedback and the introduction of multiple tries.

It is also encouraging and noteworthy to see that some of the level 2 students recognised that the system offered a useful aid to learning, and that this number has increased with the introduction of the new feature in 2004.

Conclusions

There is evidence to suggest that the improvement of the analysis of student responses to numeric problems, the consequent improvement of the relevance of the feedback provided and the introduction of multiple tries at a question, have been successful in encouraging the engagement of students and has provided a useful learning experience. There is also evidence that some students are aware of this and appreciate it.

An important body of work that needs to be carried out before further improvements are introduced is to reconstruct the experiences of each student who gained tries from the scheme (particularly those who failed to resolve their problems) so that the feedback given for each set of circumstances can be observed. This should reveal shortcomings in the specificity of the feedback, whether due to inappropriate assumptions made on the part of the tutor or due to lack of precision in the analysis of the students' responses. In either case it will indicate how the system should be improved.

Another way of improving the overall system would be to address the problem of what a student is meant to do when the feedback and materials the feedback points to, prove insufficient. In this case the student could be helped by turning to a real tutor. A communication system thus needs to built into the system so that the problem can be discussed on-line in a technically appropriate way.

A combination of a testing system able to offer multiple question tries with numerical (or symbolic) expression analysis of the type described, with a learning environment able to provide supporting information in a structured way and having a communication system suitable for live-tutor support would, it is conjectured, be able to support high quality learning in a variety of numerate disciplines, even at a distance. This is the goal toward which the SToMP project is currently striving.

References

Bacon R A "Assessing the use of a new QTI assessment tool within Physics" CAA 2003.

Bacon R A "Responding to student expectations for assessments" CAA 2004

IMS V1.2.1 IMS Question and Test Interoperability. Version 1.2.1 Final Specification <http://www.imsglobal.org/question/index.cfm> (March 2003)

Dick Bacon and Graham Smith "Adaptation of Computer based Assessment schemes to meet student expectations". ALT-C 2003 Conf. Sheffield. September 2003.

Porter, M.F., "An algorithm for suffix stripping", Program, 14(3) :130-137 1980

Arthur W Chickering and Zelda F Gamson. "Seven Principles for Good Practice in Undergraduate Education" AAHE Bulletin, 39(7), 3-7 1987

Neil Strickland, "Alice Interactive Mathematics", MSOR Connections 2, 27 2002