[O7] To assess or not to assess: what’s in a question?

Dick Bacon
Department of Physics, University of Surrey
r.bacon@surrey.ac.uk

Keywords: e-assessment, feedback, adaptive assessment, synthesised tutoring

Abstract

That modern students are over-assessed is not really in doubt so why, then, are people like myself working hard at apparently introducing more assessment in the guise of e-assessment. The answer lies partly in nomenclature - can any learning be achieved without questions being asked? - and partly in intent - are we asking questions to show how clever we are, to find out how little the students know, to provide a portfolio of achievement, or to provoke the students into thinking about their subject? This paper will explore two of the ways in which, so called, e-assessment techniques can be applied to improve students’ learning experiences without adding to the assessment burden.

In the sciences we want our students to be able to apply themselves to problems that require several steps to get from question to solution. This type of problem can be very time consuming and demoralising for those students who cannot ‘see’ the whole solution, and who, therefore, do not know what it is that they need to know to be able to solve it. One approach to overcoming this problem is to give students practice at a variety of such problems within an environment that will guide them if they need it. One such environment is a tutorial, but these are expensive. Another such environment can be created electronically, and is described in this paper. Items are expensive to develop but the costs can be ameliorated by collaboration, so that each item is used by many students.

One feature of electronic assessment systems that is probably not used to best effect is the immediacy of the feedback that is possible. The design, application and implementation of this immediate feedback to best support learners and their learning will also be discussed.

Introduction

The work patterns of science students have undergone considerable change over the last few decades, with an ever increasing accent on the assessment of components of courses rather than of the whole course. One consequence is the reduction in use of the relatively benign coursework assignments that used to be an integral part of courses, providing opportunity for consolidation of new material acquired in lectures and its assimilation into the framework of the subject. Students’ difficulties with such coursework could be handled in tutorials or problem classes, where feedback could be provided at the time it was needed, i.e. when the problem was being addressed.

With the increased pressures on academic staff time and the decrease in use of such coursework, the opportunities for students to receive timely feedback on misconceptions and misunderstandings are becoming fewer. When feedback on submitted coursework is provided, it is usually after a significant delay, and is thus of considerably reduced value to the student. To be of maximum use, constructive feedback needs to be immediate, and the student needs to be in an environment in which they are encouraged to use the feedback immediately to update their concepts and to apply them, probably in the solution to the problem that revealed the misconception.

One way of resurrecting this sort of environment is to use carefully designed questions that provide graded and constructive feedback, and then to present these to students in an electronic assessment system that provides opportunities for the students to apply the feedback immediately, either by re-trying the original question or by moving to the next step in the development of a more complex problem. The technology to create such learning aides has been available for some years, but the investment in academic time to develop such resources can be very large. There are two areas of recent development, however, that can help spread the cost of such work. One is the development of new international standards for the interchange of electronic questions between assessment systems. The other is the development of software tools for the development of such learning materials through the use of assessment systems.

In the past, tutoring systems were based around presentation systems that were able to ask simple questions. What is needed in the sciences is the ability to ask more sophisticated questions and to have more sophisticated analytical tools for the responses. New systems being developed are able to support such features, and there is promise that in the near future there will be a variety of such systems on which the materials will be able to be implemented with little change.

Enabling developments

There have been a number of developments over the last few years that have improved the prospects of new uses of assessment systems along the lines described in this paper.

New specifications have been developed for the interoperability of learning materials (or Learning Objects) between various presentation systems. These include aspects such as accessibility issues, the metadata used to describe the materials and therefore the means of searching, content management, learner information, questions and tests. This last specification, for questions and tests, is now in its second version, and has already provided a stimulus for a number of commercial systems to support the export and import of questions and assessments in a common format. Interestingly, it has had the effect of provoking some academic developments that have been direct implementations of the specification. The specification, or course, had to be sufficiently flexible to allow most of the commercial systems' question types to be represented. The academic systems, however, tend not only to match the flexibility of the specification so that they are able to render the exported questions, but are also relatively easy to extend to support new features.

The two examples below are both based upon such an assessment system that has been extended beyond the provisions of the Question Test Interoperability (QTI) specification that it implemented, which was QTI version 1.2. This specification has, however, has been superseded by a new version (QTI v2.1) that in fact supports some of the new features described, so that much of the functionality necessary to achieve the sort of learning resources being described here will be accessible as new assessment systems based upon the most recent specification become available.

The new version of the Question Test Interoperability specification provides the means for the randomisation of numeric and textual values, for creating adaptive questions that change according to the users' responses, and for retrieving values from user responses and using them to create alternative answers for later parts of a question (error propagation). These offer exciting opportunities for the development of a new class of interactive materials that create dynamic learning opportunities for students. These materials may also come to provide a means of low profile assessment of individual student progress that more closely matches that of the subjective assessment of each student carried out by a tutor during a tutorial.

There is no doubt that the development of these sorts of materials is a demanding and time consuming task, and only justified on a large scale if the resulting materials can be used widely. The interoperability specifications are, of course, of crucial importance here because the task is too large for individual departments, and collaboration between departments depends upon the ability to move the materials into the local assessment system without re-implementation. It also provides the means to retain the same materials during the evolution of assessment systems within a department.

Another maturing technology that is important to this materials-sharing scenario is the electronic repository or item bank, where these assessment items can be shared between the contributing sites. The searching of item banks requires consistent and high quality metadata, and if the items are to be shared in more than one repository then the metadata must also comply with a suitable specification, of which a number exist. Each item within an item bank can require more than one file, of course (e.g. a question file and an image), and so there is another specification dealing with the aggregation of these files into one container (called packaging, with the container being a .zip file). These specifications (learning object metadata, packaging, QTI) are all fairly recent, being developed internationally with contributions from UK HE via the JISC CETIS activity.

Example 1 - problem tutor

This resource is aimed at the type of numerical problem in which a student is given data from which a final answer must be obtained by selecting, possibly deriving, and applying appropriate formulae. This sort of problem occurs frequently in scientists' professional life, and they make good exercises for students. The selection of the appropriate formulae helps remind the student of the meaning and significance of the terms and what they represent, and the merging of different formulae into a desired form for a particular problem can help resolve misconceptions about the concepts to which they relate.

When a new topic has been introduced, it is conventional to show students one or two examples of such problems being solved before giving them their own to solve. In some cases students need tutorial or peer help to enable them to understand the concepts sufficiently to solve such a new problem at all. An adaptive assessment system can be used to synthesise a tutor's contribution to this learning activity. The problem needs to be broken down into steps, and then at each step a number of options given to the user - some appropriate and some inappropriate. The inappropriate options should be good 'distracters', i.e. should not be able to be logically deduced as 'wrong' without understanding the problem, and should include popular misconceptions so that users can be disabused of these.

Another task for the author is to decide how far to follow options that are not 'wrong' but merely 'inappropriate' or just 'inefficient'. If two vector forces are to be combined, for example, the student might be offered the options of taking moments or resolving the forces into components. Whilst the preferred method is to resolve the forces into components, taking moments (with care) turns out to be equivalent. The decision that has to be made by the author is whether to just explain that it is equivalent, to demonstrate that it is equivalent, or to provide the additional materials to let the student work through the alternative method and find out for themselves. This latter approach can lead to confusion if the alternative method is much more complex than the preferred method.

Example 2 - rich feedback

The tutoring scheme described above can be very useful in helping students develop their approach to problems, and in giving them confidence in their ability to solve them. The nature of the tutor is such, however, that a student will always eventually be able to solve each problem, even if they take a lot of false turnings in doing so. There is therefore a requirement for assessments in which the student has to solve the whole problem without help, and then submit an answer for marking.

A dissatisfaction that the students have expressed with such assessments is that having solved a problem, worked out and entered an answer (or whatever action is required for the user to indicate the answer) that the impersonality of the lack of immediate feedback about the correctness or otherwise of the answer is somehow more distressing than simply handing in the equivalent answer on paper. From the tutor's perspective, of course, feedback at the submission stage is inappropriate for non-invigilated coursework assessments where students may be working together. This leads to a situation where students spend far too long worrying about whether a particular answer is correct or not before submitting. They also dislike the inability to submit textual working, and what they therefore see as the all-or-nothing marking used. Students prefer to be able to submit working along with a final answer, knowing that they will get marks for correct working, even if the final answer is incorrect.

The submission and marking of such working by an electronic assessment systems is a difficult problem, even the automated marking of essay-type free text entry questions is not yet routinely available. Trying to understand the rationale behind an incorrect argument based on symbolic and numeric expressions is sometimes beyond the ability of the tutor, let alone an e-assessment system. This problem would become much simpler, of course, if a structure was introduced into each problem, but that would preclude the testing of the very skill that the students are being required to develop.

In an attempt to improve this situation, three numerical questions within an assessment involving several question types were modified to allow more than one attempt each. The numerical questions were selected for this use because the numerical values were randomised so that each student (probably) received a different version of the question. Therefore, giving feedback about a result would not directly help a second student without their understanding how the first student had undertaken the problem. The scheme allowed up to four tries at each question with reduced marks available for each subsequent try, and with more detailed hints being provided at each stage. This information was given to the student in an introduction to the whole assessment, and in an explanatory paragraph in the feedback when each further try was about to be given. The assessment system being used (the SToMP system) supplies feedback as each question is submitted (as opposed to at the end of the whole assessment).

Two advantages were anticipated, 1) the submission of the answer would no longer be seen as such a final action, and 2) the feedback could be applied to the problem immediately, whilst the problem was still fresh in the mind of the user.

It is clearly important that the feedback given to the student should be as pertinent to their error as possible, and not just a generic description of how to solve the problem. The feedback needs to be based as closely as possible upon the evidence the student has supplied in their response, and therefore a number of new test schemes were implemented.

- The user's response was checked against alternative answers based upon popular errors for each problem. For example, where the probability of the success of a trial was to be calculated, the probability of failure (its complement) was checked.

- The user's response was checked to see if it was in error by a simple power of ten. This can be the result of a scaling error in putting the final value into the required units.

- Where the final value was calculated from one or more summative terms involving values to be scaled, then a scheme was devised where errors in these scalings could be detected.

- Users were additionally invited to enter a numerical expression involving the original values given in the question, scaling values and physical constants. If the nature of any error in the final value could not be determined, then this expression was analysed for

    - its value (compared to the entered value)

    - the presence of the original question values

    - the use of the original question values and other constants in the required form (e.g. A = B / C)

Each of the three numerical questions remodelled in this way ended up with ten or more recognisable ways of getting the answer wrong, and during the ensuing trial of the system with 36 students there were a total of 177 attempts at these questions, with only 33 responses not being recognised by one the above schemes.

Outcomes

The problem tutor has the capability of recording all student actions and the time at which they were taken. This produces a large amount of data that, for any one tutored problem, is interesting but of limited use. For example, the rate at which a student progresses successfully through the problem can be deduced, but this merely shows that one student's working habits are different from another's. Parameters such as the ratio of correct to incorrect (inappropriate, etc.) responses, the rate of responding, the number of alternative answers chosen per unit time (can indicate guessing) and others can be extracted easily, but comparisons between students reveals little that was not already known. The real use of this information, it is conjectured, would be to log the change in some of these parameters for the same student with similar examples over an extended time scale. This information is not yet available, but could perhaps be used to create metrics of a student's academic progress.

The problem tutor has been used with a short pre and post exposure on-line questionnaire that was last year improved by the addition of a comment facility. The comments obtained so far have all been positive, one such being:

> 'This system is very, very good. I think I really enjoy in the process of learning the stuff that I thought it was quite boring, when I was using the computer based tutoring system. I hope we can use it as often as possible!!!'

The aim of the 'multiple-try with improved specificity of feedback' scheme was to turn what was purely an assessment into a learning experience. In its original form, in 2003, the most difficult of the three questions involved had 2/38 users get full marks, 9/38 obtained a reduced mark (e.g. an inaccurate value) and the rest got no marks at all. In 2004 students were invited to enter an expression, but it was only evaluated rather than analysed. In practice no student gained from entering the expression and in fact only one student obtained any mark at all. The system described above was implemented for the 2005 cohort of 25 students. For this question 21 were offered more than one try and of these, nine got it right on the second try, two on the third try and one on the last try. Only three students failed to get any mark.

In the comments of the survey of this assessment, instead of 89% of students disliking the lack of marks for working (as in 2003), only 19% cited this but 30% made some positive comment about the expression entry and multiple try scheme. With this clear recognition of the change by students, and the difference in the number of students being able to complete the question, it is difficult to avoid the conclusion that some learning had taken place, and that the resulting mark was a better indication of each student's understanding and ability for this question than in previous years.