RESEARCH DIRECTIONS

Generating Large Question Banks of Graded Questions with Tailored Feedback and its Effect on Student Performance

Stephen H. Ashworth

University of East Anglia, UK

Abstract

Formative assessments have been developed to give students practice with simple mathematical manipulations necessary in physical chemistry. A series of computer programs has been used to generate large pools of questions to accompany a core undergraduate physical chemistry module. In each assessment the questions are graded to increase the cognitive load gradually and feedback is tailored to each individual response. Furthermore, the assessments are arranged in a "daisy chain" to ensure that one is completed before the next may be attempted. The difference in terminal examination results between cohorts prior and post intervention is presented. There appears to have been a positive effect on the examination performance of the post intervention students.

Keywords: question banks, graded questions, formative assessment, feedback

Introduction

There are currently many sources of questions for computer-administered assessments. Question banks accompany many textbooks and are made available to adopters. Others have been generated as a resource by projects such as the LTSN (Adams *et al.* 2002, Bacon & Chin 2008). While these are often useful for occasional questions they tend not to be suitable for deployment as large-scale formative exercises. because the range of problems tends to be limited and is unlikely to be graded.

Corresponding author:

Stephen H. Ashworth, School of Chemistry, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, UK
Email: S.Ashworth@uea.ac.uk

In addition, the feedback is usually rather generic. To overcome these limitations a series of computer programs has been used to automate the generation of large pools of similar questions. The process is designed to enable feedback to be tailored at the level of an individual response. For example, a student who does a calculation using the ideal gas equation and omits to take into account the number of moles of gas could receive the following: "Your answer is correct for one mole of gas but in this problem you were dealing with 2.4 moles so the answer should have been. . .". If each question in a pool deals with a different number of moles, this tailored response gives the appropriate number in each separate case.

In a given assessment each question is drawn from a pool of 50 to 100 questions of exactly the same form. In order to increase cognitive load slowly the problems within a topic are graded on moving from pool to pool. The first question in the assessment is drawn from the first pool and might require straightforward substitution of values in an equation. This might be identified with Sweller's means-ends-production type 1 (Sweller 1988). The problem has a specified goal and the equation is (or should be) known, so the only unknown is the goal itself. The next question in the assessment on the same topic is drawn from the next pool. In this case a unit conversion may be required. The addition of the sub- goal means that the problem may be identified with Sweller's means-ends-production type 2. In addition the distractors are designed to ensure that common mistakes, such as lack of unit conversion, can be picked up and the tailored feedback designed to point this out.

The primary goal was to provide a resource that would direct students to practise simple mathematical manipulations which are essential to fully understand chemical concepts. However, the approach meshes surprisingly well with the best practice in formative assessment and feedback. This has been distilled into seven tenets by Nicol and Macfarlane (2006), and at least five of the seven are implemented in this approach.

Taking their points in order, good feedback practice

1. helps clarify what good performance is: the exercises help the students understand what level of performance is expected of them;
2. facilitates the development of self-assessment in learning: the feedback is specific to a given response. This will help the engaged student to reflect on what might have been missed;
3. delivers high quality information to students about their learning: as the students must continually engage with these exercises, and can revisit them as often as required, the specific feedback to a given response means they obtain high quality and immediate information on their learning;
4. promotes teacher and peer dialogue around learning: unless students work together it is unlikely that peer dialogue will develop. They are not discouraged from this but equally are not actively encouraged to work together.
5. encourages positive motivational beliefs and self-esteem: these assessment tasks are inherently low-stakes. Studies suggest that motivation and self-esteem are likely to be enhanced by the implementation of many such low-stakes assessment tasks with the feedback reflecting achievement and progress;
6. provides opportunities to close the gap between current and desired performance: these opportunities are available as the students may practise as often as required and the assessments will stay available until the students have graduated;
7. provides information to teachers that can be used to help shape teaching: as yet no information from the formative assessments themselves has been used to do so.

The goal of these assessments is to help the students with the incremental steps of problem solving rather than to develop full problem solving techniques. It is for this reason that details of full solution steps are not required. Hence fading, such as outlined by Renkl *et al.* (2004) is not implemented.

At the University of East Anglia the preferred Virtual Learning Environment (VLE) is Blackboard 9.1, and Respondus 4.0 is available for question authoring. The question pools were developed using Excel spreadsheets and Respondus then administered through the Blackboard VLE to accompany the teaching on a core first year physical chemistry course, "Energetics and Spectroscopy".

Implementation

The ideal situation would be to generate large sets of questions directly in the chosen virtual learning environment. Coding each question individually allows tailored feedback to be given for each response. However, this approach is far too time-intensive for large question pools. In this case we are dealing with over 3,000 individual questions. An alternative is to use a formula to generate a large number of questions but, at least in Blackboard, the responses would have to be formulaic and generic if coded directly in the Blackboard VLE.

Often a more convenient alternative is to generate questions in a specialised program (such as

doi:10.11120/ndir.2013.00006

Respondus). Such programs are designed to produce paper exams in addition to electronic versions. The questions can subsequently be imported into the chosen VLE, either directly or by means of a suitably packaged zip file. Equally, a program such as Respondus may be capable of importing a file with a well-defined structure in a suitable file format. This opens up the possibility of using Excel, or a similar spreadsheet program, to generate a suitable text file. Hence we can adopt the strategies outlined in Figure 1. The second strategy in Figure 1 is necessary when text replacement in the original file is required. When generating questions that involve scientific notation, Excel outputs numbers in the form 1.23E+04 rather than 1.23×10^4. Having made the required replacements the file may be imported to Respondus, where it can be checked and exported to a format ready for BlackBoard. There are a number of question types which can be treated this way. Each has a mechanism to indicate the correct answer and pairs of columns containing responses and the corresponding feedback.

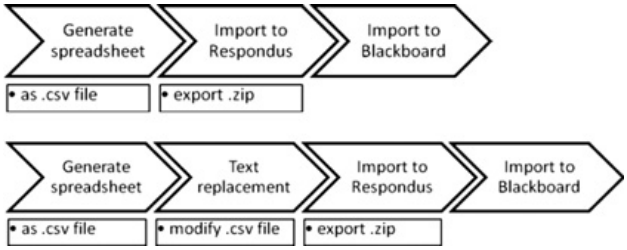

Figure 1 Outline strategy for pool generation. Respondus and Blackboard have been used as illustrative examples. The strategy will work with alternative programs. (a) the simplest strategy; (b) the modified strategy.

This approach allows one the opportunity to include calculated values in the feedback. For example, when asked to do a calculation the feedback response for the answer might be "Your answer would have been correct for one 532 nm photon. The question asks for the energy in 1 mole of photons, so the correct answer is. . .". It is possible to arrange the spreadsheet so that all the intermediate steps in the calculation are available to be used in the feedback, so that appropriate parts can be picked out and used in the construction of feedback phrases that are unique to the response given by the student.

Having generated one question in this fashion it is a simple matter in Excel to fill many cells with duplicate questions. The use of data tables and random elements means that a large number of questions may be generated without duplication. In the current work pools have been produced which have been arbitrarily limited to fifty or one hundred questions. In principle these are only limited by choice or ingenuity in the amount of randomness built into the questions.

As the primary goal was to increase familiarity and confidence with mathematical manipulations the majority of questions dealt with calculations. A few, however, were designed using snippets of sentences to put together random questions testing understanding of concepts, such as irreversible and reversible processes in thermodynamics.

The questions were graded to increase cognitive load gradually. The first pool of questions, used for question one in the test, might have all the quantities for a calculation given in units so that they may be used directly. The second pool of questions may have volumes quoted in dm^3 rather than m^3 or temperature in Celsius rather than Kelvin, thus necessitating a unit conversion. The third pool of questions might require a unit conversion and in addition the calculation of the number of moles of gas, and so on. Although these have been implemented as formative exercises each question was given a point value, from one to three. The point value of the question gives an indication of the difficulty and hence cognitive load. These marks were only used to avoid the problem of the VLE recognising an "attempt" as a student simply opening the test, but not engaging with it. The pass mark was set such that at least one of the three point questions had to be answered correctly to "pass" the test.

Each test was arranged in a chain so that test number two was only accessible to those students who had "passed" test number one. Students were able to take the tests as often as they liked and to use them for revision if required. The carrot to encourage the students to engage with this chain of formative tests was that only those who had passed up to and including test number four would be able to take the first summative test, and those who had passed tests up to and including number seven would be able to access the second summative test.

There were several drivers that encouraged the implementation of this approach. This method was implemented in a first year module and one aim was to try to foster a spirit of engagement with formative course material, especially that which is supplied on-line. The main goal, however, was to ensure the students practised routine mathematical manipulations in order to try to reduce barriers to understanding chemical concepts. To determine whether the intervention was indeed effective, the students' assessment scores over the past three years may be examined. The intervention has been

in place for two years; these scores may be compared with the previous year's scores. This approach was first implemented in the academic year 2010/11 and continued in 2011/12. From 2010/11 the course tests were necessarily administered via the VLE and students were thus at liberty to take the test at remote locations. As a result the course test became an open book format so the only results which are comparable are the results of the terminal examination which is taken in the May-June exam period. The exam results were first checked to ensure that the distribution was approximately normal. This was done using a Q-Q plot comparing the exams in each year to a normal distribution with the same mean and standard deviation. As can be seen from Figure 2, there is only a small deviation from the normal distribution (a deviation from the line x = y) which occurs at the lowest and highest marks. This small deviation is to be expected as the assumption that the marks are normally distributed implies that there should be a finite, albeit small, possibility of marks over 100% and under 0%. Given that the deviation from normality is small, the exam marks have been shown in Figure 3 as normal curves. The vertical lines indicate twice the standard deviation from the mean and therefore enclose 95% of the area under the normal curve. This plot shows that there was a significant improvement between 2009/10 and 2010/11 but no significant change between 2010/11 and 2011/12. This lends weight to the hypothesis that these formative exercises have had a beneficial effect.

There are, of course, confounding factors. One of which is that the cohorts are not the same students. Judging the students' prior ability is not straightforward even with strict before and after testing. Admissions qualifications, however, are available and may be used as a proxy for students'

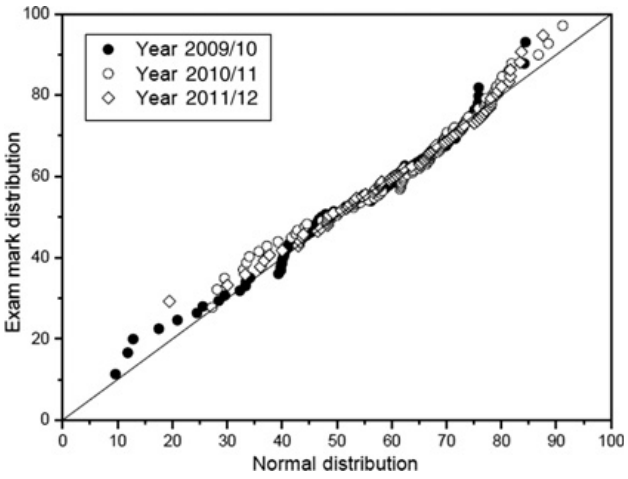

Figure 2 Q-Q plot of exam results against a normal distribution with the same mean and standard deviation.

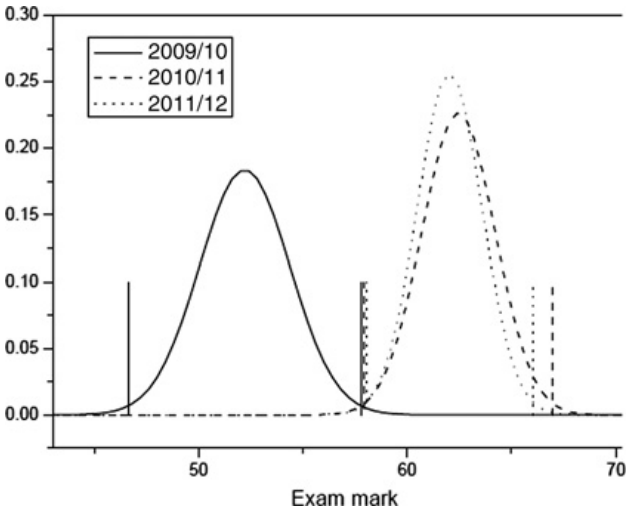

Figure 3 Exam marks from 2009/10, 2010/11 and 2011/12 for the core physical chemistry module at level 1.

prior ability. Figure 4 shows the UCAS tariff points used for admission, namely the best three qualifications including chemistry but excluding general studies and critical thinking. The introduction of the A* grade first had an effect on the 2010 entry so the tariff for A* has been counted as the same as an A grade in making this comparison.

The distributions show that the cohorts admitted to chemistry have rather different prior qualification profiles in the three years under consideration. In the first year the majority of the cohort had a rather low tariff score which might have some bearing on the correspondingly low score in the exam compared to the later years. On the other hand, the 2011/12 cohort scores marginally, but not significantly, lower than the 10/11 cohort in the

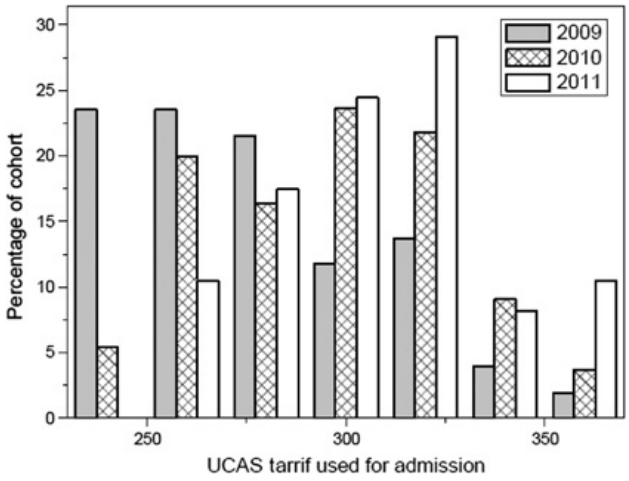

Figure 4 UCAS tariff points of the three cohorts in this study. Only the tariff of the A-levels used for admissions are shown. A* grades have been counted as A to allow direct comparison.

exam. The distribution of tariff is, however, clearly higher for the 2011/12 cohort and if the only factor were prior qualification (as measured by UCAS tariff), one would expect a greater differential between years 2010/11 and 2011/12 than is seen on the exam score. The final confounding factor is the structure of the exam itself. Unfortunately, the exam format was changed between 2009/10 and 2010/11. At this point it was brought in line with the other core chemistry first year exams with a multiple choice part and a choice of three out of five longer answer questions. Previously the longer answer questions had only been a choice of three out of four. It may be that the additional choice in the exam has also had an influence on the change in the exam marks.

Conclusions

Looking at the exam scores alone, one is tempted to conclude that the intervention described here has significantly increased the scores of the 2010/11 and 2011/12 cohorts. The significance of the conclusion is only slightly weakened when taking the prior ability of each cohort into account, but weakened still further given the change in the exam format between the 2009/10 and the 2010/11 examination periods. Nevertheless, taken together there is evidence to support the assertion that the formative assessment regime has supported the students' learning and enabled them to perform better in the terminal examinations.

It would be instructive to canvass these cohorts now to see whether the mandatory nature of the intervention has had a protracted effect on their learning or their attitude to formative work. It would also be instructive to obtain the students' opinions as to whether similar question pools would be useful to them in the second and subsequent years of their course.

Acknowledgements

This article is a summary of an oral contribution to the Variety in Chemistry Education conference 2012. SHA would like to thank M.Seery most sincerely for his encouragement and comments on the draft manuscript.

References

Adams, K., Byers, W., Cole, R. and Ruddick, J. (2002) in *LTSN Physical Sciences Development Project*, LTSN Physical Sciences Centre, http://www.heacademy.ac.uk/assets/ps/documents/downloads/25 downloads/computer_aided_assessment_in_Chemistry_1.pdf.

Bacon, D. and Chin, P. (2008) *Distributed e-Learning II Programme Final Report*, JISC/Academy Subject Centres.

Nicol, D.J. and MacFarlane-Dick, D. (2006) Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education* **31**, 199-218.

Renkl, A., Atkinson, R.K. and Grosse, C.S. (2004) How fading worked solution steps works - A cognitive load perspective. *Instructional Science* **32**, 59–82.

Sweller, J. (1988) Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science* **12**, 257–285.