ON-LINE FORMATIVE ASSESSMENT ITEM BANKING AND LEARNING SUPPORT

Sarah Maughan, David Peet And Alan Willmott

On-line Formative Assessment Item Banking and Learning Support

Sarah Maughan Group Manager, Curriculum and Teacher Support
David Peet New Technologies Manager
Alan Willmott Awards Development Manager
Cambridge International Examinations
(UCLES/CIE)
1 Hills Road
Cambridge
CB1 2EU
UK

Abstract

Access to the Internet now makes it possible to deliver new services to centres around the world including on-line formative assessments for use in the classroom. The results of these assessments provide information that can feed back into the learning and teaching process to highlight where improvements can be made. This can be a productive tool in improving student learning and has significant potential to provide a richer educational experience. For this potential to be developed, large item banks containing questions with known operating characteristics are required so that valid and reliable assessments can be built. Questions can be stored as assessing particular learning outcomes, levels of attainment, skills or other features, thus allowing specific feedback to students and to their teachers indicating curriculum areas or skills in which students were relatively strong or weak.

The limitations on the types of questions that can be asked on-line and marked objectively by computer limits the use that can be made of the results of on-line assessment. As a greater variation in the types of questions becomes available, so the use that may be made of the results increases.

As item banks are used to build assessments for known cohorts of students and results are collated over a period of time it becomes possible to supply more meaningful feedback to the users of assessments. The use of calibrated banks, and careful data management will extend this use.

The future for the Cambridge on-line assessments will be determined by the opinions of the teachers as to which forms of feedback are the most useful for themselves and their students.

Keywords

Item banking; on-line assessment; formative assessment; teacher support; item calibration; measurement; syllabus support.

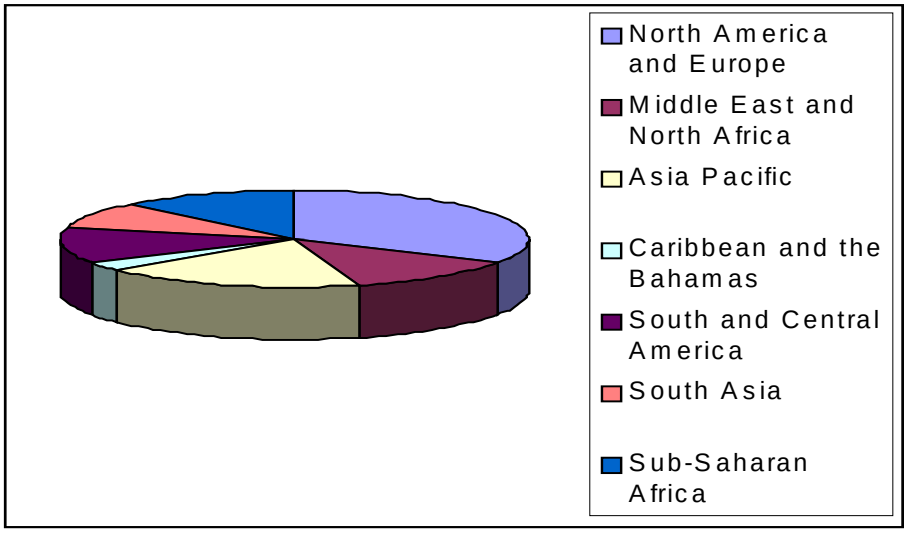
On-Line Formative Assessment

The University of Cambridge Local Examinations Syndicate provides assessments for two million candidates in over 150 countries each year. The traditional provision has consisted of an assessment syllabus followed by a related pencil and paper based summative assessment after two or more years of teaching.

Access to the Internet now makes it possible for CIE to deliver new services to centres around the world. One such service is the provision of on-line formative assessments for use in the classroom linked to an on-line Scheme of Work for the relevant syllabus.

Work by those such as Black and Wiliam indicates that there is an increasing interest in the interaction between learning and assessment. Some final examinations may be perceived to be only weakly linked to what students actually learn, but effective use of formative assessment 'can lead to significant learning gains' (Black and Wiliam, 1998). Good formative assessment can help to improve practice by emphasising personal learning, by encouraging the acquisition of higher-level skills and by focusing on criterion-referenced rather than norm-referenced performance.

In this pilot project, we took a novel approach to the delivery of materials to our Centres by not only providing guidance on how to deliver the course but also in terms of providing low-stakes intermediate assessment. The project took place between July 2000 and March 2001. This required a significant amount of resource in-house as well as considerable contributions from teachers and examiners. The site was launched to teachers on 1 December 2000, with about 150 teachers in 33 centres participating by invitation. These were distributed across the different regions in which UCLES operates.



There were a significant number of centres in sub-Saharan Africa which declined to take part, owing to the poor telecommunications infrastructure.

Materials were provided for seven of UCLES' IGCSE syllabuses. These are international versions of UK 16+ examinations. The subjects were chosen under criteria of large entry (cost-effectiveness and access to a large number of centres), availability of resources and difference in content and teaching style. The latter was important to test the concept in a variety of contexts.

Although the detailed provision varied between subjects, the key elements were the Scheme of Work and the assessments. These were delivered on two separate servers, with appropriate hyperlinks between them.

The Schemes of Work were written by examiners with a wide range of teaching experience in this age range. The syllabus was broken down on pedagogical grounds into about ten teaching units, each therefore constituting roughly half a term's work.

Alongside each unit two tests were constructed.

In most subjects, about 400 items were commissioned against a grid, with item writers asked to address specific content, skills and level of difficulty. The overarching criterion was that the items had to be able to be computer marked, which may have limited the range of skills which could be assessed to basic knowledge and understanding. As Milligan suggests, 'objective type testing is often too simplistic a measure of a learner's grasp of the course materials' (Milligan, 1998). However, in this project we used experienced item writers to produce the items, thus ensuring that an adequate range of skills could be assessed. It was very apparent that the validity of the tests (in terms of their construction) was a clear function of the definitions we produced for item writers and test constructors. The time spent on these specifications was time well spent.

All the items were then reviewed for accuracy and style at a residential meeting, before consultants (all subject specialists) constructed the tests according to an agreed protocol. One version of the test was made available in 'learning mode' where the candidate could attempt a question a number of times and also reveal the correct answer. The other test, with the same items, was set in examination mode, where although the candidate could go back and change an answer, no additional help was available.

The pilot project used the CUE assessment engine, developed at Heriot-Watt University with a contribution to funding from UCLES (see http://www.calm.hw.ac.uk/cue.html). The assessment software provides the expected functionality of an XML-based item editor, test construction tool and a server system. Using a variety of presentational formats, the question types available include straight multiple choice, multiple select, word match and hotspot. An unusual feature is a 'steps' function which allows, at the candidate's decision, a question to be presented in a more structured way. This system renders mathematical expressions well and one question type allows different forms of

algebraic expressions to be identified as equivalent as well as the incorporation of random variables within prescribed limits. The images associated with the questions were saved in .gif format.
The marking algorithms are quite sophisticated, allowing strings of characters to be forbidden or classed as obligatory. The length of the answer string can also be specified; spelling errors can also be accommodated.

Alongside the 1500 pages of content for the seven subjects, the tests comprised around 800 items with 200 diagrams.

Clearly the uptake of the tests was dependent on their relevance to the existing teaching programmes in place in centres at the time. It would have been unrealistic to expect widespread adoption of these pilot materials in the timescale available. The activity logs show clear evidence of teachers trialling some of the tests themselves and then using them with their classes. The ways in which this work will be extended are described below, but it is appropriate to highlight some of the issues that have emerged during the evaluation phase.

It is clear that it will still be some while until centres are equipped with hardware in sufficient quantity to allow large-scale testing to take place, particularly if the tests are not adaptive. Further, the quality of the Internet link is clearly a significant issue in many areas and we will seek to install a delivery system that runs on a local server in the next phase of the project. We would also hope that this system would be able to deliver a paper-based version of the same questions. There were also many of a less technical nature issues raised by teachers that will also need to be addressed.

Item Banking

The potential of the effective use of on-line formative assessment will only be realised when large item banks containing many different types of questions are available. To operate successfully in a subject area it is instructive to consider the scale of the assessment resource that might be necessary. For example, a given subject could readily be divided into ten major areas of study and each area into five sub-areas. If there are five topics in each sub-area, then, with 20 questions per topic (to allow for questions over a range of content and difficulty), then this one item bank will need some 5000 questions. A bank of this size may turn out to be only just sufficient for use with groups of students on any regular basis, as it may often be difficult to fulfil test blueprint requirements unless there are more items.

In order to be able to recall questions from an Item Bank for use, either individually or as part of a test as required, questions need to be classified under a number of headings. The level of specification will depend to a large extent on the sophistication of the assessments and the nature of the reporting systems operated and, thus, on the specification of the test blueprint. Headings such as content, educational objective assessed and difficulty are clearly needed. So too is information on how to avoid similar

items appearing in the same test and avoiding a mix of questions where one question provides the answer for another. Information on the previous use of an item and its history of use also helps to build up a detailed picture of question use that can be made available to the users of a bank.

If questions are to be calibrated, then they will all need to be pre-tested on a range of appropriate students so that substantial question response data are available for analysis. These data would, ideally, contain responses from a wide range of potential students so that detailed question statistics could be derived. At the same time, it will be possible to look at the quality of measurement provided by items and issues such as the relative difficulty of different types of questions and any bias between different types of students can be investigated. For the analysis of question data the one-parameter Item Response Theory (IRT) model, the Rasch model, will be used. This will allow various hypotheses to be investigated during the analysis while providing questions with 'known' operating characteristics. (see, for example, Rasch, 1960; Wright and Stone, 1979)

It is unlikely that the data collection for bank building could reasonably be carried out in a single step or in a very short period of time as both the production of questions and the availability of sufficient students are likely to limit the speed of progress. Accordingly, a detailed Item Bank Development Plan will be needed to allow the bank to be grown as required. Such a plan would enable a core of questions to be pre-tested on a body of students and allow for an analysis that would provide the basis of a calibrated item bank. Then, pre-testing of further sets of questions would be conducted but the new questions would be interspersed with questions already in the nascent bank. This re-use of existing questions would allow for both a check on their calibration and also for new questions to be calibrated on the same measurement scale as those already in the bank. In this way the bank is able to grow as new questions are added.

It is quickly apparent that to establish a viable item bank for a given subject area (i.e. one that is sufficiently large and calibrated) is far from easy. It will take substantial resources delivered over some period of time and will also be consequentially expensive. In HE/FE, the number of staff available for writing questions and the groups of students available for pre-testing, is often limited. In such circumstances, a mechanism for sharing questions between institutions becomes immediately attractive, if not inevitable, if a calibrated bank is to be used for assessment (formative or summative). The adoption of tight item specifications/ classifications for use in question writing will help the exchange of questions between institutions as the nature of what is being assessed will be that much clearer. This can only aid the item bank development procedure. The exchange of questions will still not be wholly straightforward but, as was clear from the pilot work, it would be a great step in the right direction.

When an item bank with calibrated questions is available, tests can be built that provide results that can be reported in terms of a common, previously identified scale. This scale would have been identified during calibration. It may also be possible to say something about 'unexpected responses' from students where responses are given that

appear to be inconsistent with the overall performance of a student or other students taken as a whole.

With formative assessment the kind of feedback given to students (and to teachers/lecturers) is central to the usefulness of the procedures adopted. If the feedback is poor or lacking in meaning then no one will benefit from the assessment process. With good criterion-referenced feedback targeted around known standards, there is an opportunity to provide a basis for self-evaluation and learning for students. To achieve this, a substantial effort is needed to establish good questions for the item bank and to continue to collect and analyse the response data when tests are drawn from the bank. Despite this effort, however, the benefits are also considerable with self-assessment hopefully leading to improvement in learning.

Calibrated item banks allow tests to be built that are flexible and fit for purpose while providing valid and reliable assessments. Item Banking is a complex process and question statistics can be influenced by the nature of the samples of students used for calibration and by their educational background/experiences. However, as questions can be stored with many associated classifications attached this allows the reporting of specific feedback to students and to their teachers. This feedback would indicate curriculum areas or skills in which students were relatively stronger or weaker than their peers and allow them to focus their future work in areas where it would be most useful.

Within UCLES a Local Item Banking System (LIBS) has been used for some years. This system has been developed in house for use with assessments in the area of English as a Foreign Language (EFL) and currently holds about 140 000 questions. Items are held in MSWord format and have many attributes of the kind described above attached. For the future, however, the use of Word is much too restrictive on question format (features such as layout, fonts, headings, etc. are all Word specific) and a system is needed that separates question content from its presentation. A new system for holding questions will be based on XML (eX*tensible Mark-up Language*) which is a development from HTML (*Hyper Text Mark-up Language*), a well-established standard. Using XML the layout of questions when rendered on the screen can be specified quite specifically and largely independently of the mode of presentation. The design of a new XML-based Item Banking System to handle large numbers of questions and, potentially, a number of concurrent users with flexibility is currently under consideration. A key feature of such a system will be the database system to be used. A standard system such as Oracle is a likely candidate.

Learning Support

As mentioned at the start of the paper, formative assessment has been shown to improve the learning process. However, for it to be really effective a number of factors are important.

1. Information about a student's strengths and weaknesses within different aspects of the assessment must be made explicit.
2. Students must be given a clear indication of where they ought to be in each of these areas.
3. Feedback must be provided by the teacher for the student, showing how any gap between achieved performance and expected performance can be reduced.
4. Students and teachers must have access to the assessment results immediately, or with a short delay, if it is to feed back into the learning process while it is still an accurate reflection of performance.

Most on-line assessment engines currently available give students results in terms of a total mark only. It is clear from the four factors given above that for test results to be used to feed back into the learning process then they need to assess in a way that gives maximum information to teachers and students about outcomes of learning. Questions selected during the test construction process must assess the appropriate content, difficulty and skills and feedback must be given for each of these areas. For example, questions within a Physics test ought to cover both recall of information and also application of knowledge. A mark of 10 out of 20 could mean that the student achieved full marks on the recall items or full marks on the application items or some marks from each section. Where questions are stored with associated attributes that are used in the test construction process then results in the final test could be split across the attributes. A student could therefore be given results in the form of 7 out of 10 on the items assessing recall and 3 out of 10 on the items assessing application of knowledge.

An extension of this methodology would be to use the calibrated item bank discussed above. It is likely that the recall items in a Physics test would be easier, in general, than the application questions. The 7 out of 10 and 3 out of 10 results given above could therefore be the expected results in terms of the item difficulties. Where calibrated item banks are used the results could be given as a scaled score taking item difficulty into account, clearly highlighting where a student's strengths and weaknesses lie.

If item banks are used to build assessments for known cohorts of students and results are collated over a period of time then it becomes possible to supply yet more meaningful feedback to the users of assessments. With the use of calibrated banks, and careful data management, a level of performance could be reported not just on a given scale but in relation to defined groups of students taking a subject. Thus, performance could be stated as being 'equivalent to the performance of the top 10 per cent of students at this age', for example. This is something already done in a number of paper and pencil tests (see, for example, Checkpoint (at www.cie.org.uk), and CELPT, Willmott and Kam, 1990)

It is still clear, however, that the role of the teacher is central to the effective use of this information. Black and Wiliam showed that the provision of diagnostic information was not sufficient without teachers and students discussing how the areas of weakness could be improved. Similarly Gipps (1994) defined formative assessment as having a

component where it is made clear to students how the gap between achieved and expected performance can be closed.

There are current limitations on the **types** of questions that can be asked on-line and marked objectively by computer and this continues to limit the use that can be made of the results of on-line assessment. As experience grows and there is a greater variation in the types of assessment so the use that may be made of the results can become more meaningful.

Three of the main areas in which the use of on-line assessments can be extended in the short term are by using self-evaluation, on-line marking of essays or by using simulations.

In our recent pilot of on-line tests as part of the teacher support site a number of teachers queried the lack of cohesion between the on-line tests and the final examinations that the students would face. Although it is possible to assess application of knowledge and other higher order skills in objectively marked tests most assessment engines do not allow for the marking of longer, more in-depth responses to questions, essential for the valid assessment of the synthesis of a number of ideas.

A number of educational resources, both on websites and on CD-ROMs, use self-evaluation of assessment as a means of offering more in-depth assessment. For this to be most effective a prompt must be given to which the student is required to produce a response before a model answer is given. The self-evaluation must be focussed using questions which highlight particular content or skills which the student can evaluate in their own work as compared to the given answer.

An alternative form of this type of assessment would be to give real answers that were awarded marks at low, medium and high points on the scale, so that the student can compare their own work with the work at the different levels. Examiner comments about why a particular mark was awarded would also be useful here.

A more sophisticated means of assessing longer responses would be to use one of the available essay-marking tools. A number of tools are commercially available that mark either for linguistic features or for content and research shows that the correlations between two human markers and a human marker and the on-line marking tool are very similar (Dexter, internal UCLES report). Although these tools may not be sufficiently established or proven to allow them to be used for marking of high stakes assessments they may be very useful as a means of improving learner support.

On-line assessment of longer responses or of essays could be given to students and the marks produced on-line; any queries could still be referred to teachers. As with the case of objective tests discussed above, the provision of only a single mark may not be the most useful information for a student to receive and there is still the critical role of the teacher discussing with the student the implications of the results and the areas for improvement.

An alternative approach to more in-depth assessment and one that uses the capabilities of the computer more effectively is the use of simulations. Simulations are being developed that allow students to manipulate variables within an experiment to investigate the relationships between them and then this information is used to answer given questions. With appropriate marking algorithms it is possible to use this type of assessment to assess both the outcome of the work, and also the processes that are used to achieve a given outcome.

Use of this type of assessment could really add value to the information that the teacher and students gain from assessment which could again feed back into the learning process.

Whatever the outcomes of future phases of the CIE research into providing alternative forms of assessment to teachers as a means of increasing learning support, the future for the Cambridge on-line assessments will be determined by the opinions of the teachers. The role of the teacher will be an important one in the provision of formative assessment and their opinions will be essential in determining which forms of feedback are the most useful for themselves and their students.

References

Black, P and Wiliam, D 1998 Assessment and Classroom Learning in Assessment in Education, **5**, 1.

Gipps, C. (1994) Beyond Testing: Towards a Theory of Educational Assessment. London: Falmer Press.

Milligan, 1998 The role of virtual learning environments in the online delivery of staff development. http://www.icbl.hw.ac.uk/jtap-573/573r1-4.html

Rasch, G. (1960), Probabilistic Models for some Intelligence and Attainment Tests. Cpoenhagen:Danmarks Paedagogiske Institut. (Reprinted by University of Chicago Press, 1980).

Willmott, A.S. and Kam, C.A. (1990). The Ngee Ann-Oxford Computerised English Language proficiency Test (CELPT). Paper presented at the 1990 AEAMEO RELC Regional Seminar on ‘Language Testing and Language Programme Evaluation”. Summer 1990.

Wright, B.D. and Stone, M.H. (1979). Best Test Design. Chicago: MESA Press.