University of Bradford,
Pathfinder Programme,
Phase one project

Briefing paper 2 of 2

Computer Assisted Assessment - Summative Online Assessment: Lessons Learned and Recommendations

The following briefing has been prepared by one of the institutions participating in phase one of the Pathfinder Programme. The programme was led by the Higher Education Academy in partnership with the Joint Information Systems Committee. The funded initiation phase of the phase one projects ended in July 2008 with work expected to continue at an institutional level thereafter.

University of Bradford

Pathfinder Briefing Paper

Prepared for the Higher Education Academy, June 2008

Introduction

The University of Bradford's e-learning Pathfinder project "Embedding Support Processes for e-Assessment to ensure the reliable and secure large-scale implementation of CAA" (Dermo, 2008a) has led to significantly expanded use of summative computer-assisted assessment and increased security through a dedicated high-stakes server. In addition, an effective and efficient scalable system to support the large-scale implementation of computer-assisted assessment across the University has been introduced in accordance with the HEFCE Strategy for e-Learning objectives for pedagogy, curriculum design and development: "Encourage the use of technology to enable electronic assessment [and] produce and disseminate models of good e-learning practice including assessment." (HEFCE, 2005:11)

.

The JISC e-assessment glossary (JISC, 2006) defines *summative assessment* as "An assessment generally undertaken at the end of a learning activity or programme of learning which is used to make a judgment on the candidates overall achievement. A key purpose of summative assessment is to record, and often grade, the candidates' performance in relation to the stated learning objectives of the programme."

This briefing paper is intended for any instructors considering using summative computer-assisted assessment, as well as the services and administrators who might support them in this. It will consider the main lessons that have been learned throughout this project with regard to summative assessment, illustrated by relevant case studies (Dermo, 2008b). It focuses in particular on the importance of communication, providing opportunities for student practice and principled item banking.

Communication, Communication, Communication

During the Pathfinder project we learnt that many of the major challenges encountered when supporting summative high-stakes assessments come down to pure and simple communication.

Responsibility and accountability.

Summative online assessment has taken place at the University of Bradford for many years, using a diverse range of tools, including Test Pilot, Blackboard Test Manager and Questionmark Perception (version 3). At first, this was on a very small scale, with interested pioneers working alone, or supported on an ad hoc basis. However, as the demand for summative online assessment grew, and the size of student cohorts began to rise into three figures, it became increasingly clear that an organised system of support was necessary (Zakrzewski, 1999).

Organising a summative online assessment brings together a surprising number of different services, groups and individuals, both within an academic department and

centrally. Issues of quality assurance need to be dealt with by the course review panel; questions need to be designed, piloted and approved by external examiners; module catalogues and student record systems must be kept up to date; department administrators need to make sure that student lists are complete; learning technologists must ensure that online versions of the test function correctly, and that enrolments are complete; PC cluster rooms must be booked; the correct PC image must be installed; invigilators must be ready to deal with new challenges; on the big day, a whole array of people must on hand to support if necessary. So many things to remember; so many risks (Conole and Warburton, 2005).

Given that these summative assessments are high-stakes and can count towards the students' final degree, it is vital that any HE institution does everything within its powers to manage these risks, to ensure that procedures are in place, and are followed in reality. It is of the utmost importance that individual instructors know what their responsibilities are, and what they can count on other services for (e.g. the exams office, IT services, user services, educational development team etc). This is especially important when things go wrong, as inevitably they sometimes do. Who to call when a room is double booked, when there are not enough PCs, when a student cannot log in and so on. It is also important to remember that solutions found must be scalable, bearing in mind the available resources.

The following diagram (figure 1) is an attempt to describe the current system established through the Pathfinder project. The left part of the diagram deals largely with administrative matters, the central column looks at the specific duties of the instructor, and the right hand part deals with learning technology and e-learning issues. It is strongly recommended that any institution seriously thanking about setting up institutional online assessment thinks about this model and how it would apply to their own set-up.

Administrative matters include making sure that proper procedures are followed at module review panels, then being certain that all the correct bureaucratic boxes are ticked (e.g. in the module catalogue and the student record system); you can be sure that a seemingly small omission at this stage can have serious consequences further down the line, as these administrative details serve as flags for other people in the system.

The role of the instructor can vary, depending on the experience of the user and the level of support required, but the key steps that need to be followed are giving students a mock version, considering a paper version of the instructions, and being on hand on the day of the exam itself. The role of the learning technology team will also depend on the level of support required, and needs to be highly flexible and negotiable. Typically, the instructor, as subject matter expert, would be responsible for coming up with the questions, and the learning technologist would be responsible for checking that the assessment settings, permissions and schedules are all correct. If there is a large number of questions to be created, clerical staff may also be brought in to deal with text entry. It is important, however, to maintain some kind of official gatekeeper for summative assessments, just as the exams office will be used to looking after scripts for paper-based exams, there should be an independent central body carrying out a similar role for online exams.

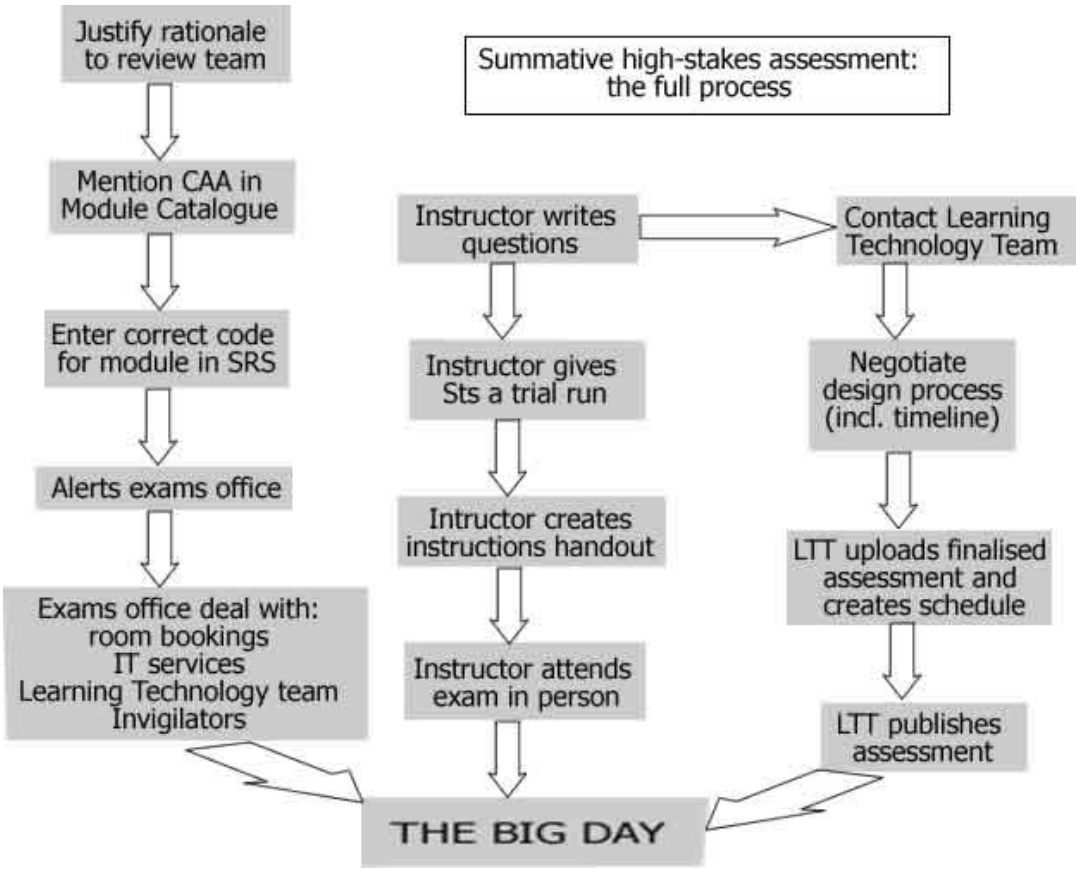

Figure 1: The Summative Assessment Process

Staff Training and Briefing

A key part of setting up the necessary support systems for summative online assessment relates to staff development (Sim et al, 2004; Warburton, 2006). At the University of Bradford, training sessions for academics are offered which deal not only with the technical aspects of designing online assessments, but also pedagogic and quality issues related to designing the kind of objective test task types that are usually found in computer assisted assessments. It ought not to be taken for granted that because an academic is an expert in their particular field of study, that they automatically know what makes a valid and reliable assessment, either on paper or online. It is also worth pointing out the fact that technology cannot magically enhance the quality of an assessment; a poorly designed question or assessment will remain ineffective, however it is delivered.

In addition, particular attention should be paid to invigilation. Many examination administration procedures were created to manage large quantities of paper, and regulations exist mainly to deal with paper-based exams. Allowing students to sit down at PCs to take examinations raises a whole raft of new issues and challenges for invigilators who may be new to the technology and who may feel that the students are considerably more technology-savvy than them. The most important thing to communicate to the invigilators is that there is a whole team of people supporting them, and they only need to focus on managing the students and watching their behaviour as before; it is not for them to worry about technical problems, log-in difficulties or problems arising from the software. It is strongly recommended, if possible, that a specialist team of CAA invigilators is recruited and trained, then there is experience to draw on and continuity from one exam period to the next.

Code of Practice and Rules and Regulations

It is vital to have a code of practice in place (Bull, J. and McKenna, C., 2004). This must be approved by the highest academic legislative body at the institution so that academics know that they need to follow procedures (just as they will be accustomed to doing with paper-based examinations). It is our experience that things tend to go wrong when individuals choose to ignore official procedures, making arrangements for themselves and neglecting to keep necessary stakeholders involved and informed. It is also worth noting that it is not enough to create the rules and regulations document and have it approved at the highest level; you must also communicate this to all interested people.

Practice Makes Perfect

The key lesson we learned here is that it is vital to give the students a practice attempt before they attempt the real exam.

Best Practice

Examinations are stressful, and it would certainly never be desirable for a student experiencing electronic assessment for the very first time in the high-stakes exam. It is good practice generally in testing to make sure that students have had some experience of any test task before the exam itself. The use of mock exams is well established in paper-based exams, but it is sometimes surprising how often people need to be reminded that these practice attempts are even more beneficial in computer-assisted assessments.

It is not necessary for the mock exam to take place under strict exam conditions, nor need the practice version be as long as the actual test. The important thing is that the students get a feel for the steps they need to go through to log in and access the test, to see roughly how the test will appear on the screen, what they need to do to answer questions and how they submit their attempt. If necessary, this mock attempt can be treated as a formative assessment, with specific feedback given. It is important that the instructions for the mock test make it clear how the practice test differs from the real summative exam.

Case Study: School of Life Sciences, Division of Archaeological, Geographical and Environmental Sciences.

The first year undergraduate module “Environmental Archaeology” has been assessed online for the past three years. This summative one-hour 50-item multiple choice assessment is delivered under exam conditions during exam week in a large PC cluster to a cohort of approximately 50 students. Since students do not have any online formative assessment in this module, it was decided to give students a practice test to get them used to the online assessment format. The practice test was shorter than the final test, consisting of just 10 questions (with 10 minutes allowed for completion) and was made available to students for one week’s duration a few weeks before the final exam as a link from the virtual learning environment, Blackboard. The instructor was able to monitor who had taken the trial exam, and send a reminder to those who

had not taken the opportunity of the mock exam. The mock exam also incorporated certain elements of formative assessment, as students were given their score and feedback, which will have helped some students in revising for the real summative exam, and gave the instructor some indication of what performance to expect in the final exam. Having followed this procedure for three years, it can be reported that there have not been any serious incidents or complaints from students in the final exam, and informal evidence suggests that they are satisfied and comfortable with the mode of assessment.

It's in the bank

The main thing we learnt from this project was not to underestimate the amount of team effort that ideally needs to go into item banking. It should also be borne in mind that the work does not end when the test has been given, as item banks need constant maintenance and evaluation using item analysis.

One criticism of summative online assessments, where students are taking the assessment simultaneously in a large PC cluster, is that it is possible for students to read answers off one another's screens. Unlike paper based exams, where the paper is lying on the table and more difficult to read, PC monitors are vertical and more clearly visible. Also, it is often the case that the PCs in these clusters are located relatively close together, which adds to the issue.

The software does, however, offer some solutions to this. First, in multiple choice questions, the order of the key (ie the correct option) and the distracters (the wrong options) can be jumbled at random; also, the order of items in the assessment can be jumbled. In this way, the chances of a student glancing at a neighbour's screen to find the answer to a question they are struggling with are very remote. Still, these solutions do not prevent cheating entirely, and yet more can be achieved using item banks. This works by building up sets of equivalent questions, and each student will receive one or more random questions from each set, so that students all receive equivalent tests, without them being identical. This offers the greatest protection against cheating – however, there are a number of considerations that must be borne in mind.

Make the categories fair

Clearly, it is vital that students receive equivalent questions. Each set of questions must be of equal difficulty; for example if you have one set with five easier questions and five more difficult, random selection of items means that some students are bound to get easier questions than others, which is not fair, and seriously affects the reliability of your test. Also, all the questions in your set must cover the same subject or topic area; for example, you cannot have five questions on geography and five on physics, as some students will receive a very different assessment, which is not fair, and seriously affects the content validity of your test. This may also affect the face validity of the test: if students feel they are not being treated fairly, their confidence in the assessment may be undermined, and you may find yourself having to justify the reliability and validity of your test. It is strongly recommended that these issues are addressed before it comes to a student making claims against the fairness of the summative examination.

The best advice we can offer from experience here is to work collaboratively. Test design is not ideally a solo task, as it is easy for an individual to overlook basic principles and fail to see weaknesses in questions they have designed themselves. It is best

practice to ask a colleague (or colleagues) to take the test first, and give feedback about content range and level of difficulty. Often, it is not possible for security reasons to trial summative assessment items as much as we would like, and we will only really get accurate data on item difficulty after the test has been given (see item analysis below).

Case Study: School of Health Studies, Interprofessional Education

For the last three years, part of the final assessment for the first-year module "Inter-professional working within health and social care" has been carried out online. This is a course that is taken by all first year students in the School of Health Studies and involves a cohort of in excess of 300 students. The 35 items for the final exam are taken from a bank of more than 80 items, divided into 12 different subject areas covered during the course. Each student is given the same number of questions from each subject set, so the subjects covered remain the same. Also, great care has been taken to make sure that the questions within each set are of an equivalent level; these questions have been used for a number of years now, and the team is confident that they are able to offer a valid and reliable item bank. A formative mock test is given, but no item bank is used for this trial, with all students seeing the same 10 questions to accustom them with the format of the exam.

Think about question weighting

Do not assume that every question has to be worth one mark. If you consider the cognitive level being tested, you may decide that some questions are more difficult or important than others (see Bloom 1956:201). For basic procedural knowledge, you might offer one mark, but may have more marks for items that assess comprehension or application of knowledge, and more yet for items that analysis or synthesis. As was mentioned above, though, items with in a set must be of the same level, so if you are going to follow this path, you must have sub-sets of question arranged by level of difficulty, and all students must receive the same number of questions from each level. See the diagram below (figure 2) for an example; there are two subject areas to be tested, and three levels of cognitive difficulty within each subject. Each student is offered a range of questions, but fairness, reliability and validity are maintained through careful design.

There is also the issue of negative marking: much has been written about this, and it is beyond the scope of this briefing paper to argue the pros and cons of this controversial topic (see Crisp, 2007 for a fuller discussion). However, it can be argued that negative marking may be used as a way of minimising the effect on cheating, especially on the questions that bear the heaviest weighting: still, the instructions must make it very clear to students how marks are being allocated.

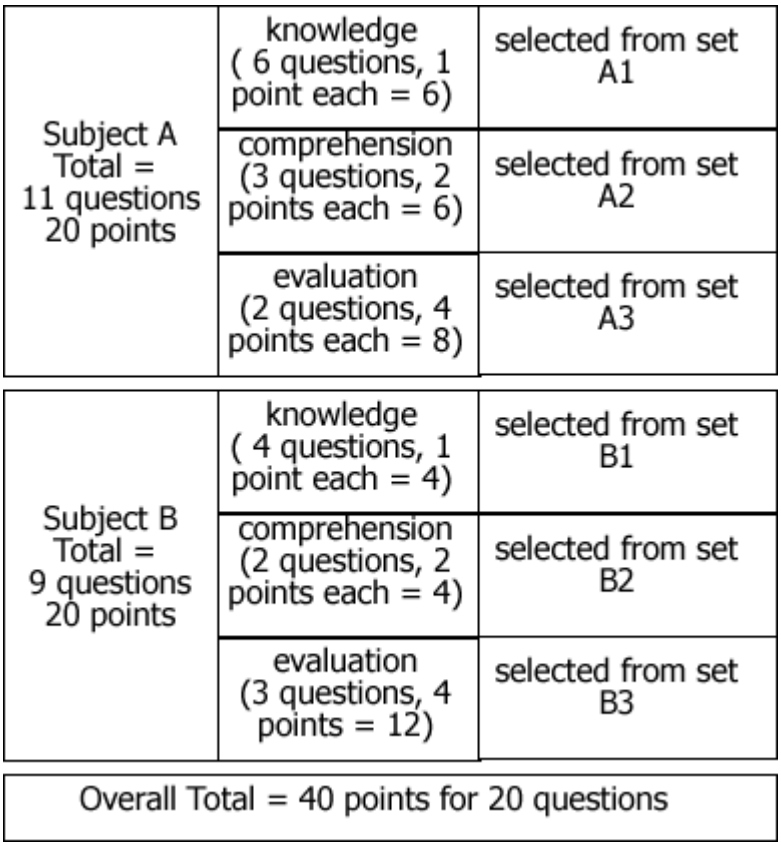

Figure 2: Item banking and weighting

Use item analysis

One clear benefit of using online assessment is that it is very easy to get item analysis data to let us look at facility and discrimination indices on individual items. This item analysis data can come in very useful if you are ever challenged on the fairness of your exam.

Basic item analysis will tell you the *facility index* of each item (ie the proportion of candidates that got it correct, ranging from 0.0 for an impossible question to 1.0 for a simple question) as well as the *discrimination index* (this compares the respective performance on this item of the bottom third and top third of the students based on overall performance on the test). The two indices will tell you how hard each item was, and whether the better students tended to get it correct. When building up item banks, you can learn a lot from this basic item analysis.

On multiple choice questions, item analysis will also tell you how incorrect answers were distributed across the distracters, which can also be very useful for editing items.

Case Study: School of Management

The first year module “Quantitative Methods in Information Management”, which is taken by approximately 200 students each year, has been assessed online for a number of years now. This is another example of random selection of approximately 45 items from item bank categories. However, in this case, as well as having ten different subject categories, these are further broken down by

cognitive level – the test has 30 “definition” questions (worth one point each), 10 “Analysis” questions (worth 2 points) and the five hardest “judgement” questions (worth 5 points each). The assessment is carefully designed to ensure that students receive the same balance of questions from each subject area, and the balance of difficulty of questions is also identical for each student. To deter students from guessing on the judgement questions, negative marking is used. Clearly, great care must be taken to ensure fairness here; students know that they receive different (but equivalent) versions, and there may always be the temptation for unsuccessful students to blame the test design and put in an official challenge, claiming that by chance they had received harder questions: this eventuality cannot be ignored and must be must be planned for. Each year, during piloting and actual use, basic item analysis is run on the assessment, indicating the facility rating and discrimination index for each question. Any question that is significantly easier or more difficult, or which does not discriminate between the stronger and weaker students, is edited or replaced. After a number of years’ use, it has been possible to fine tune the item bank.

SUBJECT A			
Question number	**Number of results**	**Proportion Correct**	**Item Discrimination**
Q01	114	0.44	0.35
Q02	137	0.64	0.59
Q03	139	0.77	0.58
Q04	151	0.75	0.52
Q05	171	0.68	0.42
Q06	117	0.62	0.49
Q07	125	0.66	0.76
Q08	170	0.29	0.16
Q09	127	0.71	0.49
Q10	111	0.95	0.03

Figure 3: Basic item analyis

For example, it can be seen in the group of ten questions above that 7 of the questions fall in the facility range .62 to .77 (ie between 62% and 77% of participants got the answer correct), which is acceptable. Three questions, however, need some attention: Q01 is rather difficult, Q08 is very difficult, and Q10 is extremely easy, and as a consequence the questions are not really discriminating. With online assessment, such item analysis can be carried out in a matter of seconds.

Conclusion

This paper has looked at three main issues emerging from the University of Bradford e-learning Pathfinder Project into Computer Assisted Assessment.

- First, it is important to try to be as open and clear as possible when it comes to describing whose job it is to do what when it comes to running CAAs: do not assume that someone else is going to sort something out, otherwise there is a chance that the work will remain undone.
- Second, make sure that students have a trial run or a mock version of the online exam before the big day. Exams are stressful enough without exposing the students to a new and unknown type of test in the high-stakes final exam.

- Third, think about item banking and how it might benefit your institution. But make sure you are treating your students fairly and are giving them a good range of subjects and levels of difficulty, and do take advantage of the potential of item analysis to fine tune your item bank.

References

Bloom, B.S. (1956) Taxonomy of Educational Objectives: The Classification of Educational Goals Chicago: Susan Fauer Company Inc.

Bull, J. and McKenna, C. (2004) *Blueprint for Computer-assisted Assessment.* London: Routledge.

Crisp, G. (2007) E-Assessment Handbook London: Continuum International Publishing

Conole, G. and Warburton, W. (2005) "A review of computer assisted assessment" in *ALT-J* Vol 13, No. 1. pp. 17-32,

Dermo, J. (2007) "Benefits and Obstacles: Factors Affecting the Uptake of CAA in Undergraduate Courses" in Khandia, F. (ed) *Proceedings of the 11th International Conference on Computer Assisted Assessment: Research into e-assessment*. pp 155-162

Dermo, J. (2008a) "Implementing Online Assessment: Finding the Right Path for an HE Institution" in Ladwa, A (ed) *E-Learning in HE* available online at http://www.rsc-yh.ac.uk/Documents/HEbooklet2_000.pdf (last accessed 12 June 2008)

Dermo, J. (2008b) *University of Bradford CAA Pathfinder Project*, available online at http://www.brad.ac.uk/lss/tqeg/projects/caa.php (last accessed 20 June 2008).

Higher Education Funding Council for England. (2005). *HEFCE strategy for e-learning*. London: Joint Information Systems Committee, Higher Education Academy

JISC (2007) *Effective Practice with e-assessment: an overview of technologies, policies and practice in further and higher education*, Joint Information Systems Committee. Available online at: http://www.jisc.ac.uk/media/documents/themes/elearning/effpraceassess.pdf (last accessed 21 June 2008).

JISC (2006) *JISC online glossary for e-assessment* available online at http://www.vyvhopescott.co.uk/jgloss/php/user_search0.php, (last accessed 28 June 2008).

Sim, G., Holifield, P. and Brown, M. (2004) "Implementations of computer assisted assessment: lessons from the literature." *ALT-J* Vol 12, No. 3. pp. 215-230

Warburton, B. (2006) “Quick Win or Slow Burn? Modelling UK HE CAA Uptake”. In Danson, M. (ed.) *Proceedings of the 10th CAA International Computer Assisted Assessment Conference*, 4 & 5 July 2006, Loughborough University. pp. 425-439.

Zakrzewski, S. (1999) “A structured approach to mass implementation of computer-based assessment (CBA)” in Brown, S. and Bull, J. *Computer-assisted assessment in higher education.* London: Routledge. Pp 171-178.