Delivering Online Examinations: A Case Study

Jason HOWARTH
John MESSING
Irfan ALTAS

Charles Sturt University
Wagga Wagga, AUSTRALIA

INTRODUCTION

IT academics need to keep curriculum material up to date in an ever-changing technological universe. Establishing a new subject accounts for just a fraction of the overall investment in time and money. A greater proportion of expense occurs in the maintenance cycle. This is especially true in the IT discipline, where subject content is frequently modified to reflect both academic advances and current industry focus.

One cost effective approach is to establish an industry alliance, and to include professional courseware in the curriculum. CSU has adopted this approach by offering programs in conjunction with a corporate partner, IT Masters Pty Ltd (http://www.itmasters.com.au), as well as with vendors such as Cisco and Microsoft. These programs contain both an academic and industry component. Upon completion of the program, graduates receive professional certification such as Microsoft's MCSE or Cisco's CCNA, as well as a Bachelor's or Master's level degree.

Students in these courses are geographically dispersed, located throughout Australia and overseas from Canada and the USA to the United Arab Emirates. Conducting conventional examinations under these conditions brings with it a host of challenges. Synchronizing examination times across a number of time zones becomes impossible when the time differences are more than a few hours. Students enrolled in these courses often find it difficult to balance work commitments against the need to sit an exam at a time and place nominated by the university. Significant numbers of students are required by their employers to travel on a regular basis and at short notice. A survey conducted in 2003 revealed that a considerable number of these students would prefer online testing to paper-based assessment. In 2004, the decision was taken to use a global testing and certification provider, Thompson-Prometric, to deliver online exams to students enrolled in the Industry Master's program. Under this arrangement, students book and then sit their final exam using facilities provided by Thompson-Prometric. One of the major changes to current practice has been that there is no longer a single examination time. Instead, students select a convenient examination time during a designated interval. This required changes to the style of examination so that academic credibility is maintained.

This paper is arranged as follows. Online testing and our experiences with the testing process are discussed in Sections 2 to 4. Statistics from the first set of online tests offered by CSU are documented in Section 5. Finally, students' perceptions of online exams and conclusions to this paper appear in Section 6.

ONLINE EXAMINATIONS

Online examinations are a cost-effective and popular means of assessing student knowledge. Bicanich, Slivinski and Hardwicke (1997) note that from a study of 400 vocational learners, 75% preferred online testing to paper-based assessment. Computer-based testing can also achieve significant cost-savings, by virtue of the speed at which results are analyzed and presented.

The system we describe, and which is used by CSU to perform summative assessment, is an

Internet-Based Testing (IBT) system. Using IBT, tests are offered throughout the world at testing centres provided by Prometric. One of the major advantages of choosing a global service provider is the network of such centres. Another is that these centres are staffed with administrators who are responsible for both the authentication of candidates as well as the invigilation of examinations. This overcomes the most significant hurdles to IBT. All tests are housed on a centralized IBT server and downloaded on a question-by-question basis. Testing is therefore carried out with minimal client-side technology. One major requirement though, is sufficient bandwidth to deal with data transmission to and from the testing centres.

There are two types of users who participate in the IBT process: candidates and authors. An author is responsible for writing and scoring examinations. A candidate sits these examinations. Within CSU, candidates are students enrolled in the Industry Master's program. Lecturers fulfill the author role. Each lecturer sets the online exam for his or her subject, which consists of a series of items suitable for use in the IBT system.

Items currently used by CSU are either objective-based or response-based. Objective-based items have only one correct answer and are graded by computer. Examples of these are multiple-choice and true/false questions. Response-based items require students to write their own answers, and are human-graded. Essay and completion-style questions are examples. Horton (2000) describes some disadvantages and advantages of various question types. (See Table 1.)

Table 1: Advantages and disadvantages of test question types.

| Question Type	Advantages	Disadvantages
Multiple choice	• easily marked by computer • suitable for use across a wide range of subject material • offer a familiar format to students	• students may guess the correct answer good distractors are difficult to create • poorly-written questions may test memory rather than higher-ordering thinking skills
True / False	• easy to create • can be marked by computer	• not hard to guess the correct answer • the question must be either completely true or completely false
Completion	• allows knowledge to be demonstrated in context	• can turn into a memory test • difficult to mark by computer because of wide variations in spelling
Essay	• most students are used to writing with a word processor	• responses must be human-graded

The overall process of preparing and delivering online examinations may divided into a number of identifiable and non-overlapping stages. It begins with the creation of the academic content that will form the basis of the examinations. This is discussed in detail in the next section. The second stage involves establishing the examination's structure and content on the IBT server. This is followed by putting in place administrative arrangements that will make the examination available to the correct candidates and within the specified time intervals. Candidates then sit for the examination and this is followed by the final data collection and marking phase once the examination window is closed off.

We presently use a testing window of five days. This was chosen to provide sufficient opportunities for students to select an appropriate examination time. This flexibility has been appreciated, as was borne out in the results of a survey conducted after the most recent examinations. This feature was rated the most important attribute of these examinations by students. We also found that 21% of students rescheduled their examination time during the test period for one reason or another.

Administration of examination access is carried out by use of a voucher number system. Students are provided with a unique voucher number which is their authorization key. It also links exam results back to a particular student. In addition, this voucher number is used to book a session at one of the testing centres. On the day of their booking, students bring their voucher number to the testing centre, along with a photo-id to confirm identity. Once authorized, students login to the IBT system with their voucher and commence the testing process. The voucher number is specific to the examination and 'expires' as soon as the student commences the actual examination.

Before testing starts, both information about the exam format and general information about the testing environment are made available through an online tutorial. This is not counted as part of the examination timing. We have found that, although most students in the Industry Master's program are familiar with online testing, many were concerned about the structure of the IBT exam. This is consistent with the findings other researchers (Horton, 2000). In light of this, information about the examination structure should be offered in advance. It should include details about what parts of the examination are human graded and which are computer graded; how much time is available to complete the test; whether questions have to be answered in strict sequence; and how candidates should respond in the event of unexpected events such as fire alarms in the building, internet dropping out, technical problems, for example, hardware failure. Providing a document that provides such information helps to allay student anxiety.

The delivery of the examination uses a minimal and intuitive interface such as the sample shown in Figure 1. Basic features include a clock that displays the time remaining, as well as navigation buttons to proceed to the next question, to end the test, and view an overall summary of the exam.
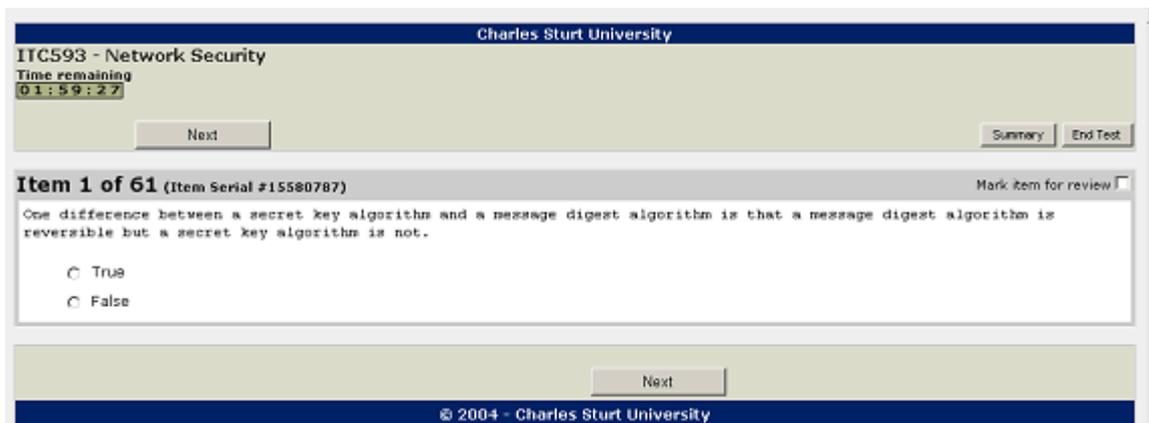
Should a candidate choose to receive a summary of the test, a listing appears of all test questions (Figure 2), including those unanswered. This is useful when a candidate wants to view questions skipped during testing. It is then possible (time permitting) for the candidate to continue and answer these unfinished questions.

Although objective test items are automatically scored by the IBT system, the candidate does not immediately receive this score. Test authors mark questions requiring human grading and once this is complete, the results are included in the overall marking scheme for the subject. At the end of semester, students receive their final grade, which incorporates all subject assessments (exams, assignments, etc), through standard CSU channels that include an electronic notification system (e-box).

Figure 2: Summary screen from the IBT system

ITEM BANKS

Each subject tested using IBT is supported by a pool of questions, that is, an item bank. Establishing these item banks is time consuming. There are several reasons for this. Devising questions that effectively test student knowledge is no small undertaking. Writing good multiple-choice questions, for example, requires creation of suitable distractors. These distractors should be plausible enough so that test-wise students cannot detect them from an obvious misfit in context (Kehoe, 1995b). The number of questions needed to populate each item bank also affects the time involved. It was decided that item banks need to be of the order of between three and five times the number of questions delivered to any candidate. This allows the item bank to be re-used several times and, with regular maintenance, has potential to be used indefinitely. In practical terms, this means that building the item bank requires more upfront work than writing a traditional examination paper, but once in place, it can achieve time savings in subsequent teaching sessions.

The use of a five day exam window means that some measures to protect the integrity of the examination need to be put in place. Item banks and an item selection strategy that involves a mixture of mandatory and randomly selected questions provides such protection. We have chosen to deliver equivalent, rather than identical, exams. Equivalence is achieved by rating questions according to difficulty and distributing content over all question categories. This means that candidates receive questions of comparable difficulty even though they are not identical.

The item types currently used in our IBT test banks are rated using a coarse classification of easy, medium or hard. In addition, individual test items are classified as either optional (selected randomly) or mandatory (delivered to all students). Students are not aware of which questions are mandatory and by making use of the editing features associated with the item banks, it is possible to select different sets of mandatory questions at different examinations. Table 2 displays a typical setup for one of our item banks.

This item bank is divided by question type and category. The total items column shows the total questions in the item bank for each question and category. Hence, for multiple-choice (medium) there are fifty items available. The optional column displays how many questions from this category should be randomly delivered to students, while the mandatory column details the number that must be delivered. Students doing a test based on this setup would receive ten questions from the multiple-choice (medium) category – nine delivered randomly from the pool of available questions (50), and one delivered to all candidates. To achieve this, each question in the item bank is flagged as either optional or mandatory.

The delivery order of the item sections, and the questions within each item section, is important. It is advisable to start with easier questions first, then to scale up the level of difficulty as the test proceeds (Horton, 2000). This builds up candidate confidence before more difficult parts of the test are undertaken. In using the Thompson-Prometric IBT system, such features are selectable within the examination structuring phase of the process.

Table 2: Setup of typical item bank.

Question Type	Category	Total Items	Optional	Mandatory
True/False	Easy	100	6	1
True/False	Medium	100	6	1
True/False	Hard	80	6	1
Multiple Choice	Easy	50	9	1
Multiple Choice	Medium	50	9	1
Multiple Choice	Hard	50	9	1
Completion	Easy	50	5	0
Completion	Medium	50	5	0
Completion	Hard	50	5	0
Essay	Easy	10	2	0
Essay	Medium	10	1	0
Essay	Hard	10	1	0

It is also worth noting that, while the first round of IBT exams included only basic question types (multiple-choice, etc), scope exists to introduce more exotic items, such as simulation-style questions or ones that involve audio responses from candidates. The technology also allows for variations of these basic question types as well as some totally new question types. A typical list would include:

- Fill in the blank
- Drag and drop
- Matching
- Multiple choice single response
- Multiple choice multiple response
- Ranking question
- Short text response
- Extended text response (even to the point of full essay)

Within that list there are opportunities to embed a variety of media types including text, graphics, audio and video. Some systems also interface the examination delivery system with other software allowing the examiner to create a style of question that is not provided for in the basic system. One interesting use of these new question styles is in the form of simulations where the student's actions are logged and assessed against a set of criteria. For example, a student might be asked to configure some IT equipment like a router and be given a multimedia representation in the form of a picture of a router, cables etc. As well as 'plugging in the components' using drag and drop simulations, switches might have to be set using buttons, and software settings altered via standard interface windows. This represents a complex task and one that has a very high level of cognitive demand. The inclusion of questions of this style greatly enhances the validity of the examination. While questions like this are not currently part of the item banks, they are planned for future implementations. A conscious decision was taken to establish the procedures and processes for conducting IBT using only basic question types that had very low bandwidth demands to 'iron out any bugs' before attempting more exotic question types.

SOFTWARE

The testing system that has been provided by Thompson-Prometric includes facilities to establish item banks as well as the administrative data need to conduct the examinations. However, this was written with a variety of clients in mind and being extremely comprehensive, was found to include a great deal of redundant data that was not required for our purposes. That, together with unacceptable time delays that were encountered during trials, led to the decision to build the item banks locally and then once completed, import them in a single upload to the IBT server. To help instructors develop their item banks, we built a web-based system that allows test questions to be entered into a series of web forms. Once entered, the questions are stored in a secure database and later, merged with the various additional data required by the Thompson-Prometric system, converted to XML and uploaded to the IBT server. A sample screenshot from this application appears in Figure 3.

Figure 3: Screenshot from CSU's item bank application.

This application allows questions to be entered and edited. There are options to specify the rating of the question, whether or not the question is mandatory, and the total marks available. The question number is generated automatically. The window on the right contains a summary of all questions entered to date, displayed as hyperlinks. The instructor selects one of these to edit an existing question.

Questions were written within a set of guiding principles such as those described by Horton (2000). To ensure these principles were adhered to, questions written for the IBT system were reviewed by a panel of instructors. These instructors also review the difficulty ratings for each question to make sure each is appropriate.

EXAMINATION STATISTICS

A well-recognised benefit of online testing is the ability to harvest performance statistics (Hopper, 1998). A detailed performance analysis can be built from this data. This may be used to pinpoint the strengths and weaknesses of students in the subject area. Statistics may also be used to identify ineffective test questions. A multiple-choice item with a distractor that is never chosen is not very effective in discriminating between good and poor students (Kehoe, 1995a). This process of test construction, review, and adjustment,

we believe, leads to better quality examinations than the traditional environment. The IBT system offers a range of statistical views to support this process. We document a few of these here.

Ranking the difficulty of items is a difficult process, especially when the item is first introduced. Long familiarity with the material may blind the author to the true difficulty of the question. The statistical tools available assist in the re-categorisation of the difficulty of items in the item bank. The following example taken from a recent IBT exam may be used to illustrate some of the issues.

Table 3: Section analysis with mean scores.

Section Name	# Deliveries	Mean Score
Fill Blank – Easy	57	56%
Fill Blank – Medium	57	23%
Fill Blank – Hard	57	40%

The results of analysis shown in Table 3 shows that questions in the medium difficulty category had a significantly lower correct response rate than those rated hard. On the surface it appears that errors were made in ranking these questions (despite each category being independently reviewed). While this serves as one indicator, there might be other factors at work. For example, students may have already been exposed to some questions in the hard category by answering similar material in mid-term assignments. Nevertheless, the availability of such reports means that the difficulty category of appropriate questions may be raised which ultimately can lead to an improvement in the quality of the examination questions and the examination as a whole.

IBT statistics provide even finer granularity than that shown above. Statistics are also available on a per-item basis. This can be useful in detecting anomalies such as the one that appears in Table 4. Question 146 was graded hard, but has been answered correctly by 95% of candidates. Question 126, in contrast, was graded as medium, but had only an 11% correct response rate. Despite the low number of deliveries for question 126, an examination of the rating of both questions is warranted.

Table 4: Item analysis with percentage of correct answers.

Question	# Category	#Deliveries	Correct
146	Multiple Choice - Hard	22	21(95%)
126	Multiple Choice – Medium	9	1(11%)

Among the many other statistical reports available we conclude with a report (see Table 5) that shows biserial correlation. In Table 5, # exams shows the number of exams that included the particular question. The correct index shows the percentage of students who answered the question correctly – that is, question 110 was answered correctly by 73% of those receiving it. Biserial correlation shows the correlation between correctly answering the item and the examinees final score. A value of 0 indicates no correlation. The data above on question 110 suggests that students who answered this question correctly tended to do well on the test overall.

Table 5: Psychometric analysis showing biserial correlation.+0.51

Question	# Exams	Correct Index	Biserial Correlation
Question_114	57	+84%	+0.51

Question_112	22	+86%	+0.55
Question_110	11	+73%	+0.68
Question_106	15	+87%	+0.26
Question_104	23	+48%	+0.26
Question_113	14	+29%	+0.1

While the discussion of various statistical analyses may or may not be of interest, the most important point is that academics in today's climate of 'doing more with less' do not normally have the resources to indulge in such exercises. Yet, the goal of providing better quality education is promoted in the mission statements of many institutions. The online system used in the conduct of the examinations for the Industry Masters courses provides these data as an integral part of the system and forms the mechanism to be able to demonstrate improvement in quality.

CONCLUSION

Reviews by students after the first round of IBT examinations were mixed. Even though steps were taken to ensure sufficient bandwidth and minimise technical problems, the total number of technical problems was unacceptably high. Some problems were not predictable and only surfaced under actual exam conditions. External events such as fire alarms during the examination, elevators on buildings not being activated due to a holiday week-end, particular keys on keyboards not working and the sudden unavailability of broadband access from the service provider were just a small selection of forces that intervened in the smooth delivery of these examinations. The pressure of using a computer under examination conditions also heightened candidates' expectations. While complaints regarding latencies (delays between completion of one question and the display of the next) were common, with 25% of candidates commenting that this was a problem they experienced, investigation of actual delays revealed only relatively small latencies of about 5 seconds in the vast majority of the questions. In circumstances of lesser pressure, such latencies might be considered satisfactory or even quite good. However, because of the stakes involved, even minor irritations featured much higher on candidates' consciousness than they might be expected to. Despite the intrusion of such negative factors, most students (82%) were still prepared to persist with the online form of examinations and saw them as the way of the future while another 15% would also like to continue their use if the technical problems are addressed.

The lessons learned in this initial implementation will be used to improve both the process and the product in successive online examinations. Making changes to workload was a key contributor to the decision to go online. While the hope of a reduction remains, the most significant aspect was a re-distribution of the associated workload over a different time frame from conventional examinations. It gives the academic a little more control over time and reduces the workload at critical periods such as immediately after examinations. More importantly however, is the belief that IBT will in time, provide a better means of achieving a quality examination process that meets modern demands of higher education such as flexibility and de-synchronisation from the 'on campus' philosophy that has pervaded most of the past. Collectively, students and instructors remain positive about the entire experiment.

REFERENCES

Bicanich, E., Slivinski, T. & Hardwicke, S. (1997). Internet-Based Testing: A Vision or Reality? T.H.E. Journal Online, September. Access date: 15 March 2004.
http://www.thejournal.com/magazine/vault/A1918.cfm

Hopper, M. (1998). Assessment in WWW-based learning systems: opportunities and challenges. Journal of Universal Computer Science, 4(4).pp. 330--348

Horton, W. (2000). Designing web-based training. New York: John Wiley and Sons.

Kehoe, J. (1995a). Basic item analysis for multiple-choice tests. Practical Assessment, Research &Evaluation, 4(10), Access date: 21 May 2004, Available at http://pareonline.net/getvn.asp?v=4&n=10

Kehoe, J. (1995b). Writing multiple-choice test items. Practical Assessment, Research & Evaluation, 4(9). Access date: 21 May 2004, Available at http://pareonline.net/getvn.asp?v=4&n=9

Contact addresses of authors:

Jason HOWARTH
Associate Lecturer
School of Information Studies, Charles Sturt University, Locked Bag 675 Wagga Wagga, NSW, 2678 , Australia.
E-mail: jhowarth@csu.edu.au
Personal web link:
http://www.csu.edu.au/faculty/sciagr/sis/home/admin/howarth.htm
Phone (office): +61 2 6933 2374 and Fax: +61 2 69 332733

John MESSING
Sub-Dean, International Information Technology
School of Information Studies, Charles Sturt University, Locked Bag 675 Wagga Wagga, NSW, 2678, Australia.
E-mail: jmessing@csu.edu.au
Personal web link:
http://www.csu.edu.au/faculty/sciagr/sis/home/admin/messing.htm
Phone (office): +61 2 6933 2418 and Fax: + 61 2 6933 2733
Interests: (Keywords or a short outline of teaching and research interests)

Irfan ALTAS
Sub-Dean, Information Technology (Industry Programs)
School of Information Studies Charles Sturt University, Wagga Wagga, New South Wales, Australia, 2678
E-mail: IAltas@csu.edu.au
Personal web link:
http://www.csu.edu.au/faculty/sciagr/sis/home/admin/altas.htm
http://csusap.csu.edu.au/~ialtas
Phone (office): +61 2 6933 2357 and Fax: +61 2 6933 2733

Interests of authors:

Jason Howarth's interests are network security, cryptography, information security, security risk management, distributed systems development.

John Messing's research and teaching interests are Multimedia/hypermedia, programming, online learning, design of electronic learning systems, computer assisted learning/teaching.

Irfan ALTAS's research and teaching interests are data mining, network security, parallel processing, numerical solution of partial differential equations, computer assisted learning/teaching.