Advisory Committee and Reviewers

Helen Ashton	Heriot-Watt University
Dick Bacon	University of Surrey
Trevor Barker	University of Hertfordshire
Andrew Boyle	Qualification and Curriculum Authority (QCA)
Clive Church	Edexcel / CETIS
Graeme Clark	Adam Smith College
Grainne Conole	Open University
Phil Davies	University of Glamorgan
Bobby Elliot	Scottish Qualifications Authority
Graham Farrell	Swinburne University of Technology, Australia
Mhairi McAlpine	Scottish Qualifications Authority
Don McKenzie	University of Derby
Chris Ricketts	University of Plymouth
John Sargeant	University of Manchester
Niall Sclater	Open University
Derek Stephens	Loughborough University
Bill Warburton	University of Southampton
Christine Ward	Ward Educational Consulting
Jeremy Williams	U21 Global
Rowin Young	University of Strathclyde

ONLINE ASSESSMENT OF LABORATORY COURSEWORK IN MICROBIOLOGY: A CASE STUDY

Malcolm Andrew

Online Assessment of Laboratory Coursework in Microbiology: A Case Study

Malcolm Andrew
The Leicester School of Pharmacy
De Montfort University
Leicester
LE1 9BH
UK
mhea@dmu.ac.uk

Abstract

This study was undertaken to examine the feasibility of assessing an undergraduate laboratory microbiology project solely online, instead of students submitting and discussing their data in an extensive written report. PHP scripts were used to construct the web forms and the data submitted by the students were stored in a MySQL database. The online assessment proved to be time-efficient for both students and tutors, albeit that the marks achieved by the cohort were higher than expected and were considered to have given a slightly optimistic assessment of the students' abilities, compared to two previous cohorts who had to submit written reports. Analysis of an anonymous student feedback questionnaire revealed that the online method of assessment was well-received by the students.

Introduction

As part of their coursework assessment in a second year Pharmaceutical Microbiology module, MPharm undergraduates are required to undertake a four-week laboratory-based project in which they perform a series of microscopical, cultural and biochemical experiments in an attempt to identify an unknown bacterial culture. After they have acquired all their data, they are expected to use an online bacterial identification program to deduce the identity of their unknown organism. Students work on their own culture and report their results individually. Traditionally, each student has had a 10 minute consultation with a tutor to verify any problematical results before presenting their data for assessment in an extensive written report. This paper describes the transition to online assessment of this project with the perceived benefits of removing the need for a consultation with the tutor, reducing the marking load for the tutor and the assessment load for the students, but without reducing the rigour of the assessment itself or the quality of feedback given to the students.

Method

The operation of the new, online, method of assessment of the project was to be a three-stage process. Stage 1 would involve the students each submitting, online, the results of 25 of their key identification tests for assessment. Upon submission, they would receive immediate online feedback stating the mark they had achieved for their results and an indication of which, if any, of their results were incorrect. Stage 2 would involve students using the existing online bacterial identification program to match their corrected results to a particular organism and so arrive at its identity. The final stage would be for each student to submit, online, for assessment, the proposed identity of their unknown culture by specifying its genus and species names.

Key requirements of a software application to handle this task were:

- a simple means of utilising electronic student data, obtained from a download file from Blackboard® - the virtual learning environment deployed by the university;

- the validation of the submitted data both in terms of the student's ID (i.e. their university 'P number') and the Code Number of their unknown culture (unique for each of the 149 students in the cohort) - in line with the QAA Code of Practice, with particular reference to assessment in flexible and distributed learning (1);

- a means of comparing each student's experimental data with the expected results for their organism and awarding a mark;

- storage of each student's submitted data, the dates of submission and the marks awarded;

- prevention of students submitting their data on more than one occasion;

- the ability to force the students to perform the tasks in the correct 3-stage order; in other words, they should not be able to submit the name of their organism (Stage 3) before they had submitted their test data (Stage 1).

Since Blackboard® itself could not fulfil all these requirements, the initial intention was to pilot the use of QuestionMark Perception® v3 as the platform for the online submission of the project for assessment. This was seen as timely since the university was intending to roll out this assessment software as a Blackboard® plug-in. Initial trials showed that QuestionMark Perception® could be adapted to do most of the above tasks, but the author found that the software was clumsy to use and that it would be difficult to upload data on students and organisms. Moreover, trialling the authored pages required their constant uploading onto the university server which was very time-consuming, particularly since much of this development work was to be done off-campus. Finally, the difficulty experienced by the university in integrating of

QuestionMark Perception® into Blackboard® led to this software being rejected for this project.

The use of Microsoft Excel® was then investigated as a means of submission and assessment of the data and spreadsheet templates were produced for this purpose. These spreadsheets would be made available for the students to record their data and submit to the tutor, by email, for assessment. The tutor would then run the Visual Basic code within Excel® that he had written to assess the data, record the mark and add comments to the spreadsheet before emailing it back to the student. Whilst this method proved to be feasible in dummy trials, in reality it would involve considerable input by the tutor to open each emailed spreadsheet, assess it using the Visual Basic code and return the spreadsheet to the student by email. Allowing macros to run on university computers to automate the procedure had unacceptable security implications. It was also uncertain whether all students were competent enough to do all this faultlessly or whether the tutor could prove or confirm the safe receipt of the students' work, in line with the QAA code of Practice (1). Accordingly, these issues led to this method also being abandoned.

The initial intention had been to use existing software to handle this online assessment to which other staff had access and which they could re-use and/or adapt for their own purposes. Despite this, the method eventually adopted involved the tutor learning to write PHP scripts to produce custom-made web page forms (2). These worked in conjunction with a MySQL 5.0 open-source database (3) running on a university web server. A direct link was then made for students to access these web page forms from within the Blackboard® website for the module.

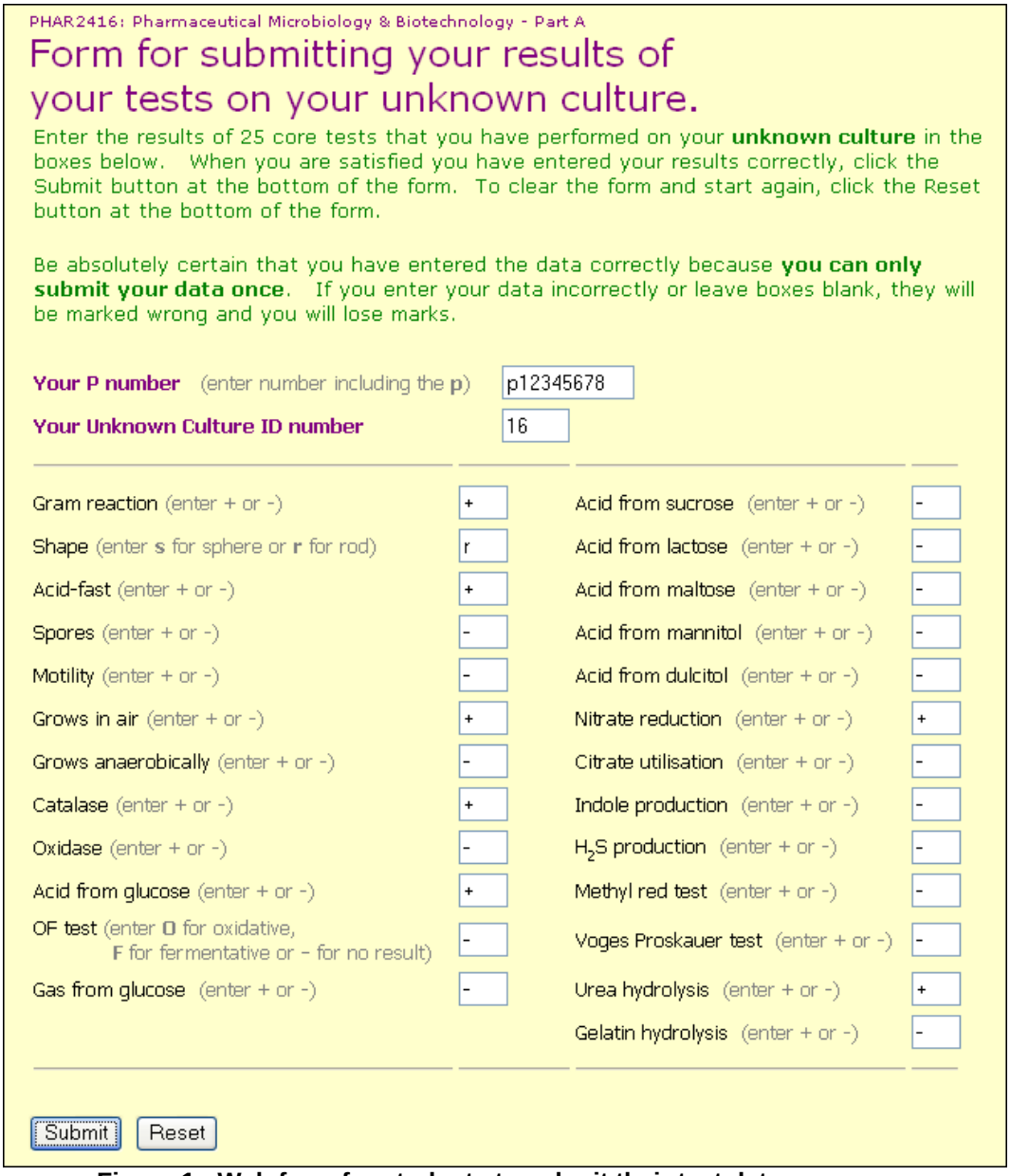
PHP (hypertext pre-processor) script is a widely-used, general-purpose, server-side scripting language that is especially suited for web development and can be embedded into HTML (2). Although it requires a server capable of running PHP scripts, there are numerous such open-source distributions freely-available on the Internet. *XAMPP* from Apache Friends (4) was the one chosen to be installed on the author's own, home computer, (running Microsoft Windows XP®) where most of the development work took place.

Fig. 1 shows part of the web form produced to enable students to submit their test results. On submission of this form, the student's university identification number (P number) was verified, together with their Culture ID Number. If either of these failed verification, or if the data had been submitted on a previous occasion, the submission was rejected and the student was given an explanation for the rejection. Otherwise, the information on the form was stored in the MySQL database, together with the submission date, and the submitted data were compared with the expected results for the bacterium concerned.

The student then immediately received a web page of feedback (Fig. 2) indicating the mark they had achieved and any of their results that were incorrect, together with comments on any problematical tests (if appropriate). They were encouraged to print off this page for future reference and told to

use their corrected data with the existing online Bacterial Identification Program (Fig. 3) to attempt to identify their unknown culture. In this program, students had to choose a column of a primary table which best fitted their data to identify their genus (e.g. Column 4, the genus *Neisseria,* in Fig. 3) then click to move to a similar secondary table to identify their species.

Once they had done this, they were required to use a second form to submit, online, a genus and species name as their tentative identification of their unknown culture (Fig.4). After verification of their personal details and a check to make sure that they had already submitted their test results but had not submitted their suggested identification on a previous occasion, these data were marked and stored in the database. They then received a further feedback web page informing them of the correct identify of their unknown culture, together with their mark for this second part of the project. Standard biological nomenclature requires the genus name to have a capital initial letter and the species name to be all lower case. If students did not conform to this convention, they were told they would receive a mark of zero for the naming of their organism.



PHAR2416: Pharmaceutical Microbiology & Biotechnology - Part A

Form for submitting your results of your tests on your unknown culture.

Enter the results of 25 core tests that you have performed on your **unknown culture** in the boxes below. When you are satisfied you have entered your results correctly, click the Submit button at the bottom of the form. To clear the form and start again, click the Reset button at the bottom of the form.

Be absolutely certain that you have entered the data correctly because **you can only submit your data once**. If you enter your data incorrectly or leave boxes blank, they will be marked wrong and you will lose marks.

Your P number (enter number including the **p**) p12345678

Your Unknown Culture ID number 16

Gram reaction (enter + or -)	+	Acid from sucrose (enter + or -)	-
Shape (enter **s** for sphere or **r** for rod)	r	Acid from lactose (enter + or -)	-
Acid-fast (enter + or -)	+	Acid from maltose (enter + or -)	-
Spores (enter + or -)	-	Acid from mannitol (enter + or -)	-
Motility (enter + or -)	-	Acid from dulcitol (enter + or -)	-
Grows in air (enter + or -)	+	Nitrate reduction (enter + or -)	+
Grows anaerobically (enter + or -)	-	Citrate utilisation (enter + or -)	-
Catalase (enter + or -)	+	Indole production (enter + or -)	-
Oxidase (enter + or -)	-	H_2S production (enter + or -)	-
Acid from glucose (enter + or -)	+	Methyl red test (enter + or -)	-
OF test (enter **O** for oxidative, **F** for fermentative or - for no result)	-	Voges Proskauer test (enter + or -)	-
Gas from glucose (enter + or -)	-	Urea hydrolysis (enter + or -)	+
		Gelatin hydrolysis (enter + or -)	-

Submit Reset

Figure 1: Web form for students to submit their test data

PHAR2416: Pharmaceutical Microbiology & Biotechnology - Part A

Review of your Bacterial ID Tests Results

Listed below are the results you submitted.
Depending on your organism there may also be a comment beside some of the tests (if appropriate).
For example, you will be told if a particular test result in the *Bacterial Identification program* is shown as '**d**'.

You should print off this page for future reference by clicking on the 'PRINT ME' link below .

PRINT ME

Your P number: p12345678
Your Culture ID Number: 16
You submitted your data on: 21-11-05

The test results you submitted were:
(Assume that each of the results you have submitted is correct, unless the word INCORRECT appears beside it)
(Tests where you have submitted no result are scored as INCORRECT)

Gram reaction: +
Shape: r
Acid-fast: +
Spores: -
Motility: -
Grows in air: + Note: your organism can grow at 52°C.
Grows anaerobically: -
Catalase: +
Oxidase: -
Acid from glucose: +
OF test: -
Gas from glucose: -
Acid from sucrose: -
Acid from lactose: -
Acid from maltose: -
Acid from mannitol: - INCORRECT
Acid from dulcitol: -
Nitrate reduction: +
Citrate utilisation: - INCORRECT
Indole production: -
H_2S production: -
Methyl red test: -
Voges Proskauer test: -
Urea hydrolysis: +
Gelatin hydrolysis: -

Your score for this part of the project is: 23 out of 25
(Please note: some of the tests carry negative marks if they are incorrect)

Figure 2: Feedback web page received after a student submits their test data

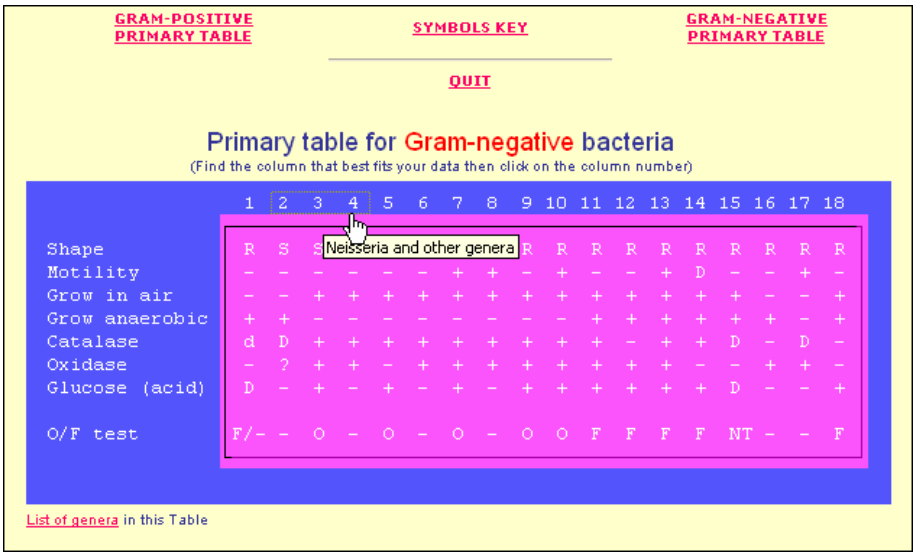


Figure 3: Part of the web page of the Bacterial Identification Program.

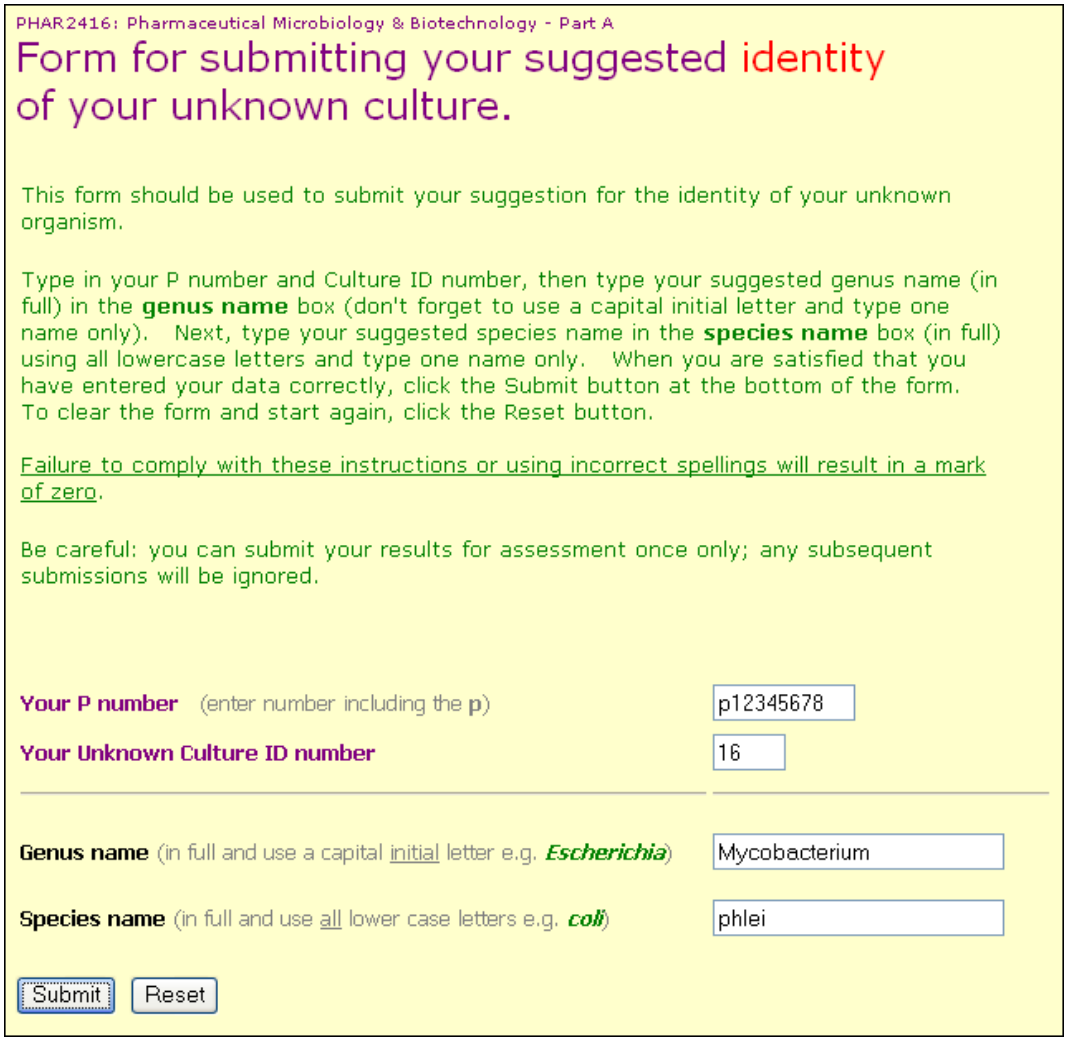


Figure 4: Web form for students to submit the identity of their unknown culture.

The change to online assessment was introduced for the 2005-2006 cohort of students. A detailed description of the nature of the assessment and the procedure the students were expected to follow was provided on the Blackboard® module website. This was augmented by a 5 minute presentation during a lecture, using screen shots similar to those used in this paper. The period allowed for online submission of the project data was three weeks from gathering all the experimental results. After this, the links to the submission forms were removed from the Blackboard® module website.

Results and Discussion

All 149 students in the 2005-2006 cohort submitted their data by the deadline, with one student completing the project within two hours of the submission forms being made available on the Blackboard® website for the module! With regard to identifying the correct organism, 133 students (89%) got the identity completely correct (i.e. both genus and species names), 6 students (4%) got just the genus name correct and 10 students (7%) failed to get either name correct. Only one student failed to conform to the convention on biological nomenclature, erroneously using a capital initial letter for the species name.

The level of attainment in this project was very high, with 83% of the student cohort obtaining at least 30 marks out of a possible 40 marks and two students obtaining full marks, so producing a skewed distribution (Fig. 5). This is far higher than in the two previous cohorts, where students had to submit a written report, instead of presenting their data online, and where the data are normally distributed (Fig. 5).

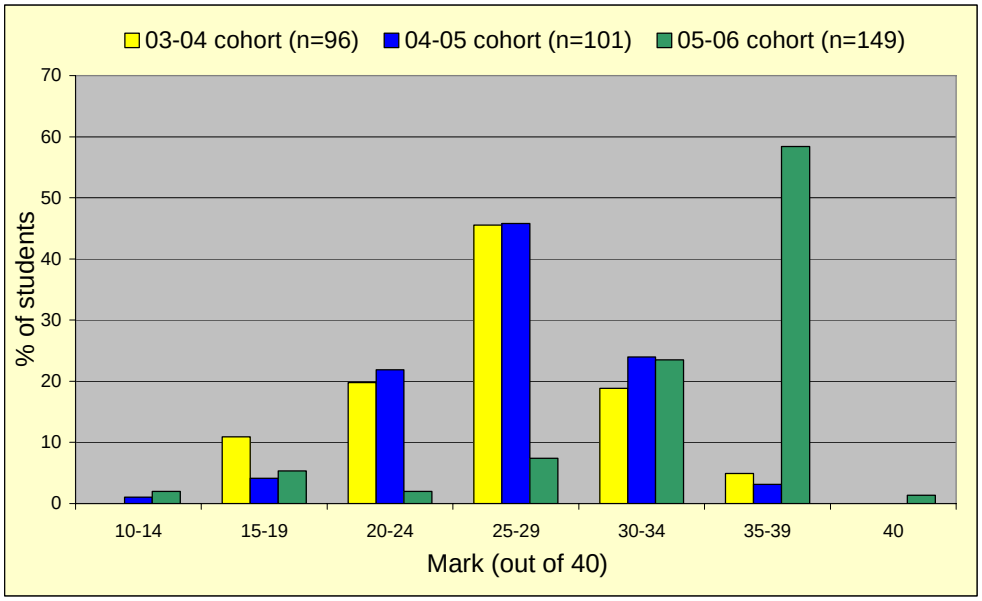


Figure 5: Comparison of bacterial ID project marks achieved by 3 cohorts of students. Only the 05-06 cohort was assessed online; the two other cohorts were assessed by submitting a written report.

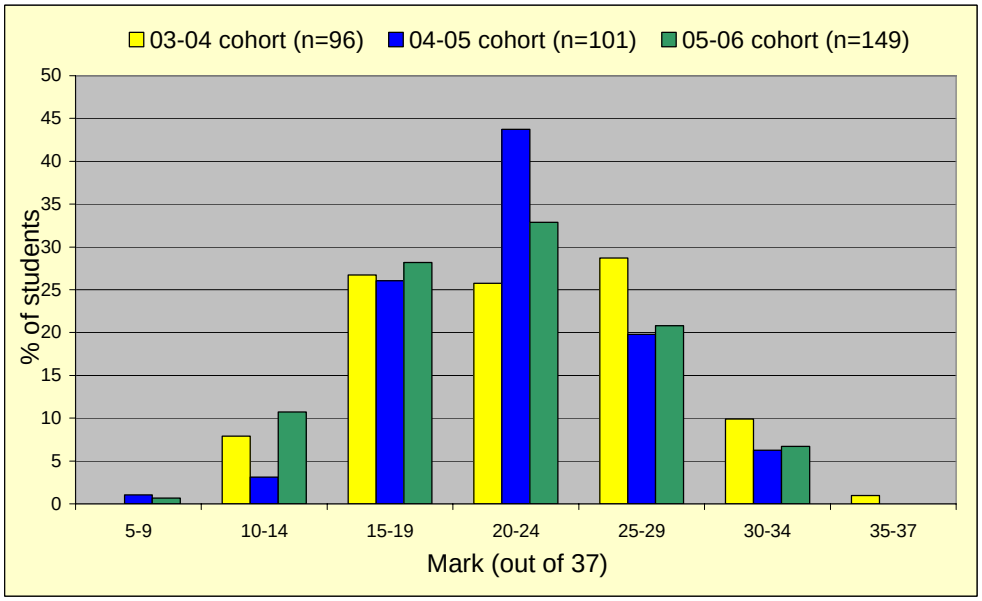


Figure 6: Comparison of non-project coursework marks by the 3 cohorts of students.

Analysis of the marks for the rest of the coursework (i.e. excluding the project), where the method of assessment had been identical for all three

cohorts, revealed that there was little difference in levels of attainment between the cohorts (Fig. 6). This suggests that the online assessment of the project had skewed performance towards an over-optimistic assessment of student ability which did not match the tutors' perceptions of the students' laboratory skills during the classes. This might be avoided in future by more carefully controlling student access to various stages of the project. It was felt that students might have been using the results of their tests to tentatively identify their culture *before* submitting their tests for assessment and submitting test results that they *expected* to obtain, rather than those that they actually obtained, to increase their mark. In future, it is proposed to prevent access to the identification program until all students have submitted their test results.

Another reason for the higher marks may have been the absence of discursive discussion, construction of tables of data and diagrams in the online assessment, which were all required, and assessed, in the written report. Today's students tend to be relatively weak in these skills. A further difference between the two forms of assessment was that the results of the tests were not specifically marked in the written report, whereas in the online assessment, they represented 50% of the project mark.

An anonymous, online, student feedback questionnaire revealed that online assessment was popular with the 05-06 student cohort (Table 1). Those students who also offered comments, generally, were very complimentary (Table 2). From the tutor's standpoint, the online assessment method will provide a substantial saving in time, now that the software has been written and the system is in place and has been tested. It will help to cope with the increasing numbers of students being recruited to this course. An added advantage is that it is a totally objective form of assessment, whereas it is difficult to avoid some subjectivity when assessing written reports. Thus, it obviates the need for double marking. In addition, there is greater scope for the analysis of student achievement by interrogating the database, which is both simple and quick to do. Archiving the database for year-on-year comparisons is also easily done.

The input of student information at the start of the project was not a huge task, mainly consisting of downloading student data from Blackboard® then copying and pasting it into the database as a text file. Recording the Culture ID Number that each student had chosen was done in the laboratory class, using an Excel® spreadsheet on the tutor's PDA. This information was then uploaded into the relevant table in the database, again as a text file.

	With regard to the Bacterial ID Project...	**Strongly agree**	**Agree**	**Neutral**	**Disagree**	**Strongly disagree**
1)	...I liked being able to present my data online for assessment	37 (54%)	27 (39%)	4 (6%)	1 (1%)	0 (0%)
2)	...I would rather have presented my data in the form of a written report	1 (1%)	2 (3%)	20 (29%)	24 (35%)	22 (32%)
3)	...I liked being able to receive online feedback on this coursework	44 (64%)	22 (32%)	2 (3%)	0 (0%)	1 (1%)
4)	...the project was a good test of my ability to analyse and interpret data	23 (33%)	38 (55%)	3 (4%)	4 (6%)	1 (1%)
5)	...the operation and assessment of the project were clearly explained on the website	29 (42%)	35 (51%)	3 (4%)	2 (3%)	0 (0%)
6)	...I found the project appropriately challenging	19 (28%)	40 (58%)	8 (12%)	0 (0%)	2 (3%)
7)	...the project was fairly assessed	25 (36%)	26 (38%)	14 (20%)	2 (3%)	2 (3%)

Table 1: Results of anonymous student questionnaire given to the 05-06 cohort about online assessment of the bacterial ID project (*n=69*).

1) I think that some people may have been at an extra advantage than others as they got a fairly easy organism and hence very good marks, and some of us had to struggle right upto the deadline.
2) I found it very interesting. enjoyed doing all the different tests etc.
3) I thought the project was really good. everyone having unknown bacteria meant we had to understand the work for urselves and it was fun. the online part was a really good idea and made everything much easier.
4) Very well organised thanks!
5) very original Dr.Andrew, liked your teaching method very much!
6) The continual application of tests to the unknown cultures we were given helped the practicals interlink with each other and gave them more purpose, since failure to complete the tests could have prevented good coursework marks.
7) Using online facilities made completeling the project more interesting and fun!
8) I found this style of assessment really refreshing.
9) Submitting online was easy an quick and enabled you to attempt the work at your own pace in your own time. However the bacterial identification program was difficult to understand, especially meanings of symbols.
10) Wonderful way of learning!
11) i didn't like the fact that you needed to get both the genera and specie correct in order to obtain the 25 marks possible. you got all 25 or nothing!
12) Excellent and interesting, thats my view on lectures, practicals and the E-learning.Its a pity that other modules do not benefit from the hardwork put in to it from the lecturers.
13) I felt that the help on Blackboard was really useful and I found it relatively easy using the ID program to identify my unknown bacteria.
14) i found this piece of coursework interesting and a good learing experience as it was new to me
15) I found this project very interesting and challenging and its was presented in a very interesting way for which I must thank Dr.Andrew, I would definately get a first class degree if all the tutors were organised like him.
16) I prefered the online report because it was quick and simple to use, which meant more time could be spent on analysing/interpreting the results obtained rather than typing out a full report.

Table 2: Student comments (literal transcription) from the questionnaire about online assessment of the bacterial ID project.

It is not intended to re-consider the use of QuestionMark Perception® or Microsoft Excel® to handle the data in future, due to the respective disadvantages of these applications (discussed above). This decision was reinforced by the ease with which a MySQL database can be maintained and re-populated, year-on-year, and the satisfaction of being able to create customised PHP web pages with little difficulty. This has led to the author using PHP scripting to create other applications, such as online student feedback questionnaires with real-time analysis of the submitted data (as used to gather the data in Tables 1 & 2).

Conclusion

It is intended to continue to assess the project online for future cohorts of students, using PHP-scripted web pages, with the attendant benefits offered by this method of assessment. However, the submission criteria and the weighting of the assessed components will be adjusted to produce, what is considered to be, a more realistic representation of students' laboratory performance.

References

1. QAA Code of Practice for the assurance of academic quality and standards in higher education
   http://www.qaa.ac.uk/academicinfrastructure/codeOfPractice/section2/default.asp#elearn

2. PHP scripting
   http://www.php.net/

3. MySQL open-source database
   http://dev.mysql.com/

4. Apache Friends *XAMPP* distribution for various platforms
   http://www.apachefriends.org/en/xampp.html

E-TESTING BASED ON SERVICE ORIENTED ARCHITECTURE

Goce Armenski and Marjan Gusev

E-Testing Based on Service Oriented Architecture

Goce Armenski
Marjan Gusev
Institute of Informatics
Faculty of Natural Sciences and Mathematics
University “Ss Ciryl and Methorious”
Arhimedova bb
PO.Box 162
Skopje
R. Macedonia
armenski@ii.edu.mk
marjan@on.net.mk

Abstract

The extensive use of technology in learning and working, is forcing its use in the assessment process. A lot of software packages exist in the market to realize automated assessment. Several of them are very comprehensive, but most of them are stand alone applications without possibilities for interoperability, adaptability according to learner characteristics and possibilities for content reuse.

In this paper we describe the purposes and the process of designing an interoperable E-Testing Framework by remodelling an existing E-Testing system and introducing new structured Service Oriented Architecture, based on encapsulating existing business functions as loosely coupled, reusable, platform-independent services which collectively realize required business objectives. This common framework should support interoperable content, exchange of data and learner profiles, and give the possibility for search and retrieval of any data bank content in local and remote repositories.

Keywords

E-Testing, E-Assessment, E-Learning, web based assessment, SOA, web services, interoperability

Introduction

The characteristics of the society in which we live, where knowledge and the ways of its use are the most important in everyday life, brings new challenges for higher education. Student-centred learning and constructivist approaches

are just some of the paradigms which have emerged, and are being supported by technological advances. Higher education institutions have the main role in the process of redefining the models for acquiring knowledge and skills. Technology is more often used in learning as a tool for lectures, delivery of materials, and assessment of student knowledge.

There are several systems for automatic assessment on the market, mainly as part of distance learning systems. However, there are independent software packages for computer based assessment, web based assessment or electronic assessment. Many of these systems are very comprehensive but most of them are stand alone applications without possibilities for interoperability, adaptability according to learner characteristics and possibilities for content reuse. [7] [6] [17]

The system for electronic testing at the University “Ss Cyril and Methodious” is known as ‘eTest’. This is the result of continuous development of concepts and software used for conducting frequent assessments. More that 500 students take part. The original idea was to create a system to help realization of exams for cases where the number of students is very large (several hundred on each exam), and in the case where each student is allowed to apply for an exam each month. This idea was later expanded to realize an intelligent and independent system for testing applicable both to conventional and distance learning.

The system is very comprehensive, catering for many aspects in the assessment of student knowledge. The basic structure of the system consists of courses for which material is divided into lectures, and organized in a tree like structure. It evaluates a different test for each student each time a test application is made, and has innovative procedures for exam strategy definition, cheating prevention, grading and results reporting. The concepts and the architecture of the system are described in [1].

In this paper, we analyse the weaknesses of the system, and propose its redesign according to the latest recommendations by the e-Learning community with a common goal to produce an interoperable, easily adaptable and more flexible system supporting pedagogical diversities.

Background

The system has been in use since 2001. Until the end of 2005, a series of 589 assessments have been realized with 9861 tests generated. The question database has 12391 questions from 26 courses. The effects of using the system were analysed in a separate study, both from a teacher and student perspective [2].

In 2005 the system was installed in 3 other Universities in the country, where it is used for assessment of student knowledge. The creation of 4 different environments posed new challenges to the research and development team.

The current system architecture does not provide complete interoperability between systems in these locations. It does not allow searching, using or modifying interoperable questions. Neither does it support the creation of joint courses where students from different universities can participate. Cross institutional cooperation through the sharing of information is required arising from the fact that many courses are beginning to be taught collaboratively, realizing the concepts of student mobility and lifelong learning.

We have decided to remodel the existing system by introducing new structured service oriented architecture. This is based on encapsulating existing business functions as loosely coupled, reusable, platform-independent services. This is to achieve better interoperability with other systems, exchange of data and content between systems using current widely adopted standards, increase flexibility and provide greater pedagogic diversity,

Service Oriented Architecture

In the past few years, world leading organizations in the e-learning community were focused on creating a joint vision for a common technical framework in the e-learning area, and in defining *international learning technology standards and specifications, in order to allow systems to "support organisational and cross-organisational processes for enabling effective e-learning*" [20]. These standards and specifications are supposed to promote interoperability, flexibility and pedagogic diversity in the e-learning process.

As a result of those activities few detailed frameworks were developed. Some of the most successful and comprehensive are:

- JISC Technical Framework to support e-Learning (ELF). [12]
- IMS Abstract Framework (IAF) [13]
- LeAPP Learning Architecture Project [14]
- Carnegie Mellon's Learning Services Architecture [5]

One common structural issue for which these organizations reached a consensus was the adoption of a Service Oriented Architecture (SOA). SOA has many advantages including reusability and flexibility of implementation, higher compatibility with the Grid, lower over all costs, protection of legacy investment, lower cost of entry, rapid development, potential for business processes to drive technology" [4].

In [21] Willson discusses the pedagogical aspects of SOA e-learning system analysing 6 pedagogical choices in e-learning, and concludes that a "Brave New World' of web-service driven environments" offers much greater pedagogical diversity than the monolithic systems.

The comparison of the above mentioned frameworks shows that they all have layered architecture consisting of a set of services which can be used in an e-learning context, and collectively realize required business objectives. The basic idea behind this is that anyone wanting to develop e-learning applications can select services, integrate them and incorporate them into the application.

SOA in Assessment

Although Assessment is present as one of the main services in all the mentioned frameworks, JISC [11] as the organisation developing the E-Learning Framework (ELF) has made significant steps forward in the definition of the Assessment domain.

Following its strategy for the creation of Reference Models for a number of domains, assessment is extensively a subject of research. Several projects have been funded [10], among which FREMA (Framework Reference Model for Assessment) is the most comprehensive. The project defined the domain, created a map of resources and "concept map of the common processes" [15], identified common usage patterns, developed use cases and defined Web Services in the domain. [9]

Another project (TENCompetence) [19] has identified assessment as a main tool for achieving its goal and have developed an assessment model based also on SOA. [18]

Modelling a Common Framework

Concentrating on remodelling the existing eTest system architecture, we have identified Web Services which will collectively realize required business objectives of our system, namely: Item Construction, Test Construction, Test Delivery, Results Collecting, Marking, Decision Making, and Statistical Analysis. Our supporting services are: Schedule, Notify and Announce, Authentication, Track, and User Management.

Recently developed frameworks and reference models are still on an abstract level and have little support in practical implementation. For example in the assessment area there isn't yet a complete product. Because of this investigation on standards and development work is underway in order to see what the results from the implementation of the proposed models will be.

A comprehensive overview of assessment projects is given in [8]. Most of them give practical realisation of particular services identified by the FREMA project, and propose extensions to (or verify) already existing standards. Some projects are more comprehensive, demonstrating the use of multiple services in SOA (ASSIS [3]).

Besides practical implementation of the proposed SOA framework in eTest, analyses of the results from its implementation, and comparison to the other

identified frameworks, our future work will pay close attention on the Test Construction Service (in FREMA identified as Author Assessment, or Assessment Construction in TENCompetence). Very few projects in the assessment area to date had analysed and discussed the problem of test construction and delivery algorithms. According to [17] there are different kinds of test delivery models.

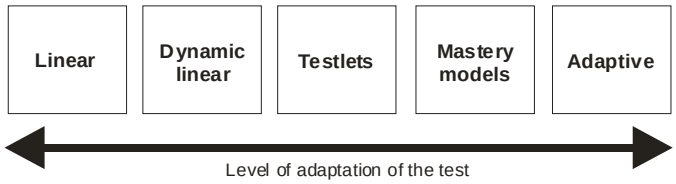


Figure 1. Models for test delivery [16]

In order to support wide pedagogical diversities, any assessment system should be able to provide all models of test delivery. By using different models for test delivery in the learning process we can simulate the world of interactive games, and hope to motivate students more than would be achieved through traditional means.

Conclusion

In this paper we have shown how our system for eTesting can be remodelled and upgraded by introducing new structured service oriented architecture.

By identifying and implementing unique functionalities, we expect to update and fulfil the existing SOA frameworks and models, creating a common framework which will provide greater interoperability, exchange of data and content, and greater pedagogical diversity.

In its practical realisation we will use the experiences and results from already developed projects in the assessment area [8]. By researching the possibilities for using different models for test delivery depending on context specifics, we expect to contribute in improvements of the diversity and quality of the learning process.

References

[1] Armenski G, Gusev M, “eTesting Infrastructure”, FACTA UNIVERSITATIS (NIS), Series: Electronics and Energetics, Volume 18, Issue No. 2, 181-204p, August 2005. http://factaee.elfak.ni.ac.yu/fu2k52/contents.html

[2] Armenski G, Gusev M, Results from using eTesting methods in CS education, Proceedings of the Workshops on Computer Science Education, Univ. Nis, 2004, pp.49-54

[3] ASSIS: Assessment and Simple Sequencing Integration Services http://www.hull.ac.uk/esig/assis.html

[4] Blinco K., Mason J., McLean N., Wilson S. (2004), Trends and issues in e-learning infrastructure development, A White Paper for alt-i-lab 2004 Prepared on behalf of DEST (Australia) and JISC-CETIS (UK).

[5] Carnegie Mellon, Learning Systems Architecture Lab. http://www.lsal.cmu.edu/lsal/expertise/technologies/learningservices/vuarchitecture.html

[6] Cristea P, Tuduce R, Automatic Generation of Exercises for Self-testing in Adaptive E-Learning Systems: Exercises on AC Circuits http://wwwis.win.tue.nl/~acristea/AAAEH05/papers/4-AIED2K5_W9_PC_RT_1.pdf

[7] Davies, W. M. and Davis, H. C. (2005) Designing Assessment Tools in a Service Oriented Architecture. In Proceedings of 1st International ELeGI Conference on Advanced Technology for Enhanced Learning BCS Electronic Workshops in Computing (eWiC), Vico Equense, (Napoli), Italy

[8] Davies, W. M., Howard, Y., Millard, D. E., Davis, H. C. and Sclater, N. (2005) Aggregating Assessment Tools in a Service Oriented Architecture. In Proceedings of 9th International CAA Conference, Loughborough

[9] Davis H, Howard Y, Millard D, Wills G, FREMA: e-Learning Framework Reference Model for Assessment http://www.jisc.ac.uk/uploaded_documents/frema_210306.ppt

[10] http://cetis.ac.uk:8080/elearningprogram/subjects/assessmentfold/assessment/topic_view

[11] http://www.elearning.ac.uk/

[12] http://www.elframework.org/

[13] IMS Global Learning Consortium: Abstract Framework http://www.imsglobal.org/af/index.html

[14] LeAP Project Case Study: Implementing Web Services in an Education Environment http://www.education.tas.gov.au/admin/ict/projects/imsdoecasestudy/LeAPProjectCaseSummary.pdf

[15] Millard, D., Howard, Y., Bailey, C., Davis, H., Gilbert, L., Jeyes, S., Price, J., Sclater, N., Sherratt, R., Tulloch, I., Wills, G., & Young3, R. (2005). Mapping the e-Learning Assessment Domain: Concept Maps for Orientation and Navigation. In Richards, G. (Ed.), Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2005 (pp. 2770-2775).

[16] Patelis, T., An overview of Computer-Based Testing, The college board, RN-09,. http://www.collegeboard.com/research/html/rn09.pdf

[17] Sclater, N., Low, B., & Barr, N. (2002, 9-10 July 2002). Interoperability with CAA: does it work in practice? Paper presented at the Proceedings of the Sixth International Computer Assisted Assessment Conference, Loughborough University.

[18] Tattersall C, Hermans H, OUNL's assessment model, January the 10th 2006 http://dspace.ou.nl/handle/1820/558

[19] TENCompetence - Building The European Network for Lifelong Competence Development http://www.tencompetence.org/

[20] Wilson S, Blinco K, Rehak D, Service-Oriented Frameworks: Modelling the infrastructure for the next generation of e- Learning Systems. July 2004. http://www.jisc.ac.uk/uploaded_documents/AltilabServiceOrientedFrameworks.pdf

[21] Wilson S, Can web service technology really help enable 'coherent diversity' in e-learning? http://www.elearning.ac.uk/features/pedagandws

BRIDGING THE GAP BETWEEN ASSESSMENT, LEARNING AND TEACHING

Helen Ashton and Ruth Thomas

Bridging the Gap Between Assessment, Teaching and Learning

Helen Ashton(1) Ruth Thomas(2)
(1) School of Mathematical and Computer Sciences
Heriot-Watt University
Riccarton
Edinburgh
EH14 4AS
(2) JeLSIM Partnership
www.jelsim.org
rct@jelsim.org

Abstract

This paper looks at how learning, teaching and assessment can become misaligned resulting in an education system that does not support student learning. It discusses a number of issues that, if addressed, could narrow the gap between teaching, learning and assessment:

- The need to use the same software tools throughout the education process
- The need to assess what the student has learnt rather than what is easy to assess
- To consider the possibility of assessing qualitatively not quantitatively
- To consider the possibility of assessing the application of knowledge, rather than its acquisition

This paper outlines how past work developing a software tool combining simulations and assessment (Thomas et al 2004, 2005) has been used to produce exemplar teaching material that uses the same software throughout the educational process. The system is capable of handling the needs of traditional e-Assessment and of providing the tools to investigate innovative assessment that focuses on performance and the quality of learning. Whilst the principles discussed are not subject specific, an exemplar in the area of Mathematics is used in this paper.

Introduction

The interconnection between assessment, teaching and learning is undeniable. As Ramsden (1992) pointed out:

"The process of assessment influences the quality of student learning in two crucial ways: it affects their approach and, if it fails to test understanding, it simultaneously permits them to pass courses while retaining the conceptions of subject matter that the teachers wish to change"

If teaching methods and assessment are not aligned to the learning activities stated in the course objectives, then a discordant teaching system results which does not support student learning (e.g. Biggs 1999).

For many years, much summative assessment has consisted of timed, closed book, pen and paper examinations. Some workers have expressed doubt about whether traditional forms of assessment are properly aligned with the teaching and learning. Biggs (1999) points out that the declarative knowledge (or knowing about things) and functioning knowledge (using declarative knowledge to solve problems) are frequently assessed in the same way and that students state what they have learned rather than show it performatively. Jonassen (2003) suggested that we should assess the performance of the learning activity rather than simply the outcome.

Other workers, such as Biggs and Collis (1982) have advocated qualitative, rather than traditional quantitative assessment, focusing more on how well students had understood a concept rather than how much they had understood. They put forward a taxonomy (Structure of the Observed Learning Outcomes, or SOLO) that was designed to distinguish between learning outcomes of low and high quality. The taxonomy is in 5 levels:

- **Pre-structural:** no understanding is demonstrated.
- **Uni-structural:** a very basic understanding with focus on one component or aspect of a complex problem only.
- **Multi-structural:** understanding of several components of a problem, but no understanding of how they relate to one another.
- **Relational:** understanding of several components in an integrated fashion so that logical conclusions can be drawn.
- **Extended abstract**: students are able to generalise their understanding into new areas and draw new general conclusions.

If there is misalignment between learning and assessment where learning does not involve ICT, does the situation improve with the increasing use of ICT to support learning? E-assessment is being gradually introduced in the domains of education and training, but it is currently at a stage where its primary function is to improve efficiency and reduce costs by duplicating the functionality and rationale of traditional assessment (Bennett 1998). Until now, the focus has been on what is easy to automate rather than what we actually wish to measure (Ridgway et al 2004) - this does not improve the alignment between learning and assessment.

Issues of alignment also arise when ICT is used in only parts of the educational process (e.g. in learning but not assessment or vice versa).

“Currently, we have bizarre assessment practices where students use ICT tools such as word processors and graphics calculators as an integral part of learning, and are then restricted to paper and pencil when their ‘knowledge’ is assessed.”

Ridgway et al (2004)

Conversely, Harding and Craven (2001) point out the difficulties of introducing ICT into summative examinations in advance of learners routinely using such software and note that:

"It is however self-evident that in order to be fair, a summative assessment system must be based on activities that are familiar to the learner."

There are a few examples where commercial ICT is used in both learning and assessment for instance, the use of Computer Algebra Systems (CAS) in the Baccalauréat Général Mathématiques examination in France and in the College Board's Advanced Placement Calculus test in the USA.

The aim of this paper is to outline the design and development of an e-Learning project that attempts to close the gap between learning, teaching and assessment and addresses the issues identified which can lead to misalignment:

- The need to use the same software tools through out the education process.
- The need to assess what the student has learnt rather than what is easy to assess.
- The possibility of assessing qualitatively not quantitatively.
- The possibility of assessing the application of knowledge, rather than its acquisition.

Whilst the principles discussed are not subject specific, an exemplar in the area of Mathematics is used in this paper. Examples of other subject areas can be seen at http://www.jelsim.org/ourwork.html, and the application of multiple interfaces is well illustrated by the Solar Transit Model at http://www.jelsim.org/content/applets/solar/index.html.

Technical Background

In 2004 Thomas et al described a fledgling system for integrating simulations and assessment in a way that allowed a range of parameters and activities from within the simulation to be monitored from within the assessment engine. The paper outlined how the system could be used to test higher order skills (classified according to Blooms revised taxonomy). The paper focused on assessment, it did not include a consideration of learning and teaching within the process. In 2005 an enhanced system was described (Thomas et al 2005) and learning and assessment were linked as the paper considered learning objectives which could be met using simulations, and ways in which these learning objectives could be assessed. A number of examples were produced (see http://www.calm.hw.ac.uk/sims-asses.html).

However, using the simulations solely for assessment effectively only considers half of the education picture. In fact the system allows the simulations to be re-used and repurposed in a number of ways throughout the education process (Thomas and Milligan 2004). It is this potential that is being explored in this project in order to address the issues outlined in introduction.

The Pilot System

The underlying tools

The system used in this pilot is an integration of the JeLSIM simulation toolkit (Thomas and Milligan 2004) with the PASS-IT assessment engine (PAE) (see http://www.calm.ac.uk/sim-asses.html) presented at the CAA conference in 2004 (Thomas et al 2004). With JeLSIM tools, a programmer creates the algorithm controlling the behaviour of the simulation (the model), and the educationalist (a non-programmer) can create the user interfaces (visualisations) to the model. One model can have many, very different, visual interfaces and different initial states. Attention can be focused on a specific concept by changing how the model is exposed to the user, and how much of the model the user can manipulate.

Choice of exemplar model

A number of new simulation algorithms (models) have been designed from scratch for this system. They have been designed to be extremely flexible as they will be required not only to fulfil current educational needs but also to permit exploration of different approaches to assessment.

In this paper we will use examples from one of these models – the curve laboratory.

The model behind the curve laboratory has a number of general features:

- a number of families of functions can be used (including straight lines, trigonometric functions and quadratics),
- controls which allow learners to manipulate the graphical representation of these functions,
- details of the expressions for the graphed functions,
- the ability to obtain information regarding the state of the model from an external source (the assessment system or learning environment),
- the ability to communicate a user determined state to an external system (i.e. communicate “answers” or user state to the assessment system or learning environment).

For this exemplar system we have used the curve laboratory and focussed in on the relationship between the expression for a trigonometric function and the associated graph. In Scotland, students usually encounter this area in Mathematics at the Scottish Qualification levels of Standard Grade / Intermediate 2 and Higher.

Often students learn about this topic by plotting functions by hand, or by comparing static graphs (as in Figure 1) to understand the relationship. Whilst it is important that students can manually create graphs of functions, creating these by hand can be tedious and may result in the general concept being lost, either due to the time taken to draw the graphs, or due to a lack of understanding about how to draw the graph. The exemplar system removes these barriers and allows the quick manipulation of various parameters to explore the relationship at a conceptual level. There is therefore the potential here to examine the approach a student takes to this exploration (be it free or

directed). In addition, the model has the potential for a high level of reuse in a number of contexts via the creation of a number of visualisations for the model, and the use of randomised and fixed initial states.

More specifically here we are concentrating on the following objectives (note: these are not always taught all at once, and not necessarily expressed in this manner). In general a student should become familiar with the relationship between changes in the expression and the graph (and vice versa) for the following manipulations:

a. changes in amplitude
b. changes in period
c. horizontal translations
d. vertical translations

In other words, students are exploring the link between the graphs and the following general expression[1],

$$a \sin(bx + c) + d$$

where the parameters a, b, c & d relate to each of the points above.

For example, a student may begin by exploring the concept of a change in amplitude, and how that relates to the parameter a in the simplified form of the above expression of $a \sin(x)$ – see Figure 1.

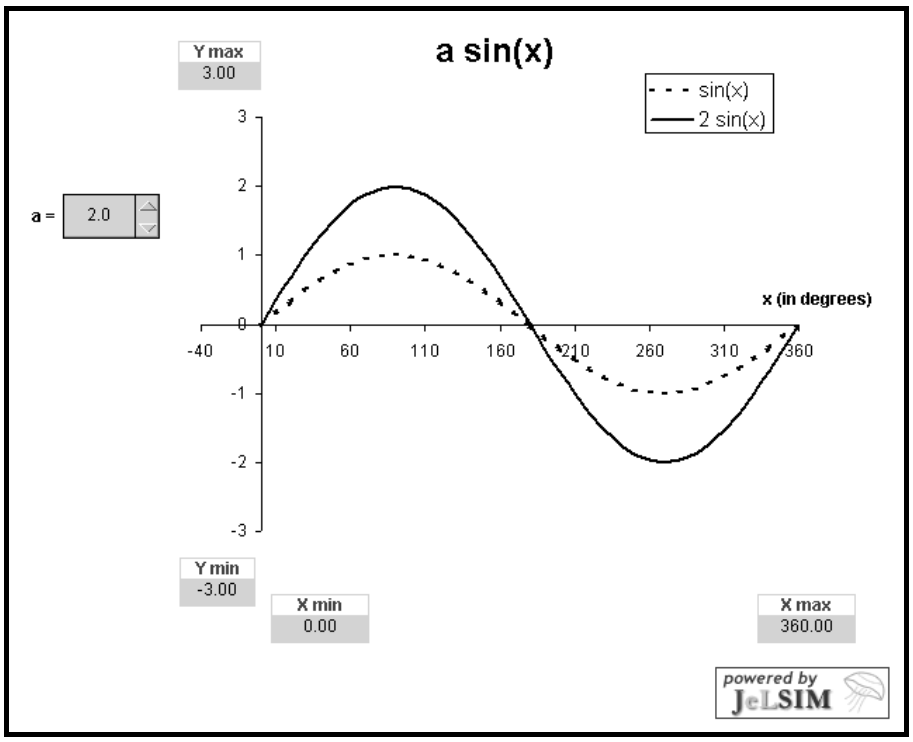


Figure 1: Graph of sin(*x*) (dashed) and 2sin(*x*) (solid)

Visualisations[2]

[1] This general expression is only one of a number of general expressions that can be utilised within the curve laboratory.

The models are used to create simulation visualisations (interfaces). Many different visualisations can be created from one model (Thomas and Milligan 2004). A key way in which visualisations can be varied is by the degree of freedom to explore that they allow the learner. This could range from no freedom where the learner sees a pre-set demonstration through a directed task, to open exploration. This means that the visualisations can be used to suit a range of teaching modes and styles. To illustrate this point some of the different types of visualisation that can be created with a simulation are listed below followed by an example of how they might be used to deal with a typical learning objective:

- **A teacher's mode** – suited for demonstrating concepts in a lecture. Such visualisations would be suitable for display on an overhead projector (using large font and highly visible colours). They contain minimal description and additional material, since a teacher will be present to provide explanation. They may also have preset functionality provided via a button click for ease of presentation.

  A teacher might start by showing a visualisation in which one parameter can be varied and demonstrate the effect of varying this parameter, getting students to predict what will happen as the parameter is changed.

- **A course material view** – suited to exposition of the subject (describing and explaining a topic) demonstrating points within online learning material.

  The course material could back up the classroom teaching by allowing the students to cover this at their own pace and in their own time. It could progressively introduce students to more variables, providing situations where the student must predict, check the feedback and adjust their own understanding.

- **Exploratory views** – suited to activities, which require the student to explore a topic. This form of visualisation can provide guided exploration where the student is prompted to undertake an activity or free exploration where the student can explore as they wish. It is possible to collect and analyse information about how the student explores the environment.

  Exploratory views are generally more open than course material. The degree of exploration (number of parameters which can be altered) can be increased as students gain expertise. This mode is designed for students to ask their own questions, make predictions to build up their own rules of how the curve and expression are related. It can also act as a "laboratory" to obtain information to solve the more complex tasks set in the problem solving view.

- **A problem setting view** – used in combination with a problem scenario or task. The simulation is set to an appropriate starting state and the student

[2] The term visualisation is used to include not only the appearance of the simulation, but also its initial state.

is asked to solve a problem. Student activity and answer generation can be linked to the assessment engine and its reporting system.

> *The problem solving view sets a task, which may be fairly complex, and provides access to the other views as resources for the students to solve the problem.*

Example course material will be available at www.calm.hw.ac.uk/CAA2006.

Creation of visualisations

An important feature of the system is that it should be capable of use by non-programming teachers. In the pilot study, the teacher is very familiar with PAE but not with JeLSIM. Teachers are not expected to create JeLSIM interfaces from scratch, a number of templates covering common question types in this subject have been created. (These question types have been selected from the SQA Higher Mathematics National Assessment Banks (NABs)). Each template has a designer's "overlay" screen, which is only visible in editing mode (not when the student runs the visualisation). The designer's overlay provides access to important variables and a tutorial in how to use them. These provide a starting point for novice designers wishing to create learning material or new assessment questions.

Combining resources to form learning material

Visualisations can be used in a variety of ways to create resources for use within a course on a subject (either as standalone objects or as components in other resources). The visualisations take the form of Java applets that can be used within any web page or learning environment.

Summary

The system as described above is capable of being used in teaching, learning and assessment, and as such has the potential to overcome the first issue identified as a problem in the introduction. In the remainder of the paper we look at how assessment, both traditional and more performance based, can be enabled by the system.

Assessment within the system

Traditional assessment

Past papers (Thomas et al 2004, 2005) have shown how traditional assessment was aided by the combined PAE-JeLSIM system where it could be used to quickly set questions and mark answers. The same approach applies when the "assessment enabled" simulations are embedded within additional learning material. The outcome of a simulation activity can be passed to the assessment engine for marking and feedback, and the student can be directed to appropriate remedial activity.

A number of questions can be grouped to produce a test. This can be used to provide e-Assessment versions of traditional paper based tests (e.g. SQA Higher NABs). The system can also be used in diagnostic assessment at a

higher level (e.g. 1st year university) to highlight and remediate weaknesses of students entering university.

The type of assessment described above is unusual in its focus on simulations and their use in assessment, but it still functions within the milieu of traditional education and its learning objectives. There are however, opportunities for learning and assessment outwith the traditional areas.

Alternative assessment techniques

As well as a traditional form of quantitative assessment, the system is being used to look at qualitative assessment and how well students make use of the knowledge they have acquired. It is useful to assess the quality of learning, not only to determine a student's mastery of a subject, but also to allow teachers to evaluate the quality of their teaching.

The SOLO taxonomy provides a convenient way of categorising the quality of learning. In terms of the type of learning expected from students using the "curve lab" this ranges from: *uni-structural*, where the student can cope with one variable, through *multi-structural* where they can handle more variables, through *relational* where they can understand the interrelationships between the variables and can easily solve problems in the domain, through to *extended abstract* where they can make generalisations to other families of curves. When the learning deepens and becomes higher quality, rather than applying rules they have been taught, it is anticipated that students will begin to build their own rules and use their own strategies to solve problems.

Can an assessment strategy be adopted to ascertain the quality of learning? The assessment must provide information about more than whether a student can determine the answer to a problem: it must look at *how* students solved the problem and the strategies they adopted.

Suppose, for example, students were shown two curves on a graph and were given the equation of one curve and asked to work out the equation of the other. They would be given access to the curve lab whilst undertaking this task. It would be of interest to know whether the students adopt a structured approach to determining how each variable affects the shape of the curve or adopt a "pot luck" approach randomly selecting variables to try. Do students engage in a form of informal self-assessment when they engage with simulations? Do they predict what will happen if they carry out an action? Do they test the result of the action and then if necessary revise their understanding? The curve lab can monitor the actions a student takes when manipulating curves and our aim is to determine whether it is possible to distinguish different types of approach from reviewing the sequence of actions undertaken by the student. Research to establish an initial baseline to link recorded actions to types of solution strategy will be carried out by (audio) recording students "thinking aloud" as they work through a problem. Asking students to "think aloud" will also be used to determine the construct validity of new forms of questions.

Often, students are given rules to allow them to solve a particular problem type, but ultimately students need to be able to develop their own

understanding and rules if they are to be able to extend their expertise to variants of that problem type. One method that will be investigated to assess this is “teach back”, a form of assessment in which the student presents their understanding of a concept to a teacher, examiner or classmate. In this case, students would be teaching their own rules for solving a particular type of problem. Within this system, this “teach back” could potentially be computer-based, as non-programmers can create visualisations and other resources. (Currently some training is required to allow a teacher to create simulation resources or assessments and some improvement of user interfaces would be required if student novices were to use the system).

Finally, students can be asked to create questions for other students. There is a considerable body of work on the benefits to learning of students creating questions (e.g. Draaijer and Boter 2005) Again this system, with its authoring capabilities, lends itself to this type of non traditional question. It is recognized that both this and “teach back” are not usually of use as teaching material, but of more use to the student. An approach such as this may also have links into other developments in the areas of learning and e-learning, such as the use of portfolios and reflective logs.

Summary

The system as described is capable of providing support throughout the educational process. Teachers can create learning material (including assessments) and students have access to learning material, assessment and feedback. It can support traditional assessment linked to learning.

The capability of the system is being further exploited to consider new approaches to assessment that enable qualitative measures to be undertaken.

References

Bennett, R. E. (1998). Reinventing Assessment: Speculations on the Future of Large Scale Educational Testing. Educational Testing Service.

Biggs, J. B. and Collis, K. F. (1982). Evaluating the Quality of Learning. The SOLO Taxonomy (Structure of the Observed Learning Outcome), Academic Press

Biggs, J. (1999). Teaching for Quality Learning at University, Buckingham: Open University Press.

Draaijer, S., Boter, J (2005). Questionbank: Computer self-supported questioning. 9th International Computer Aided Assessment Conference, Loughborough, Loughborough University

Harding, R., Craven, P. (2001). ICT in Assessment: A three legged race, Cambridge Assessment

Jonassen, D. H., Howland, J. Moore, J., Marra, R. M. (2003). Learning to solve problems with technology: a constructivist perspective., Pearson Education Inc.

Ramsden, P. (1992). Learning to Teach in Higher Education, London: Routledge.

Ridgway, J., McCusker, S. & Pead, D., (2004). Literature Review of E-Assessment. Nesta Futurelab series, NESTA Futurelab

Thomas, R., Ashton, H., Austin, B., Beevers, C., Edwards, D., Milligan, C. (2004). Assessing Higher Order Skills using Simulations, Proceedings of 8th International Computer Assisted Assessment Conference, Loughborough, U.K.

Thomas, R., Milligan, C. (2004). Putting Teachers In The Loop: Tools For Creating and Customising Simulations, Journal of Interactive Media in Education Designing and Developing for the Disciplines Special Issue (15).

Thomas, R., Ashton, H., Beevers, C., Edwards, D., Milligan, C. (2005). Cost effective use of simulations in online assessment. 9th International Computer Aided Assessment Conference, Loughborough, Loughborough University.

Project on Assessment in Scotland – using Information Technology (PASS-IT) website, http://www.pass-it.org.uk/ (accessed 1st June 2005)

Simulations and Assessment (2005): Using simulations for assessment. Available online at http://www.calm.ac.uk/sim-asses.html (accessed 1st June 2005).

MEASURING STAFF ATTITUDE TO AN AUTOMATED FEEDBACK SYSTEM BASED ON A COMPUTER ADAPTIVE TEST

Trevor Barker and Mariana Lilley

Measuring Staff Attitude to an Automated Feedback System based on a Computer Adaptive Test

Dr Trevor Barker
Mariana Lilley
Dept of Computer Science
University of Hertfordshire
Hatfield
AL10 9AB
t.1.barker@herts.ac.uk
m.lilley@herts.ac.uk

Abstract

In Higher Education today, increasing reliance is being placed upon the use of online learning and assessment systems. Often these are used to manage learning, present information and test learners in an entirely undifferentiated way, all users having exactly the same view of the system. With the development of increasingly large and complex computer applications and greater diversity in learner groups, consideration of individual differences and greater efficiency in learning and testing have become important issues in designing usable and useful applications.

Our initial findings, reported at CAA 2005, suggested that students valued this approach to providing automated feedback and considered it to be a fast, effective and reliable method. In the study presented in this paper, the attitude of staff to our automated feedback tool is presented. Three presentation sessions involving more than 80 university lecturing staff were undertaken and their views of the feedback tool were captured using video recordings. Initially a small group of computer scientists took part in a short presentation followed by a discussion where they presented their views on the CAT approach, the adaptive nature of the system and the provision of feedback. The second study involved a presentation and feedback session with more than 50 lecturers from all sectors of the university who provided their opinions of the approach in general. A short questionnaire was administered at the end of this session. The results of this, which broadly support our approach to automated feedback, are presented in this paper. A third study is reported, which involved 20 lecturers with special interests and roles in online and blended learning within the university.

Subsequent analysis of the sessions using qualitative data analysis methods showed that teachers in general were receptive to the idea of automated

feedback based on CAT. Several interesting ideas arose from the discussions, which are presented here. Computer based testing and automated feedback are becoming increasingly important in Higher Education. It is important that the views of teachers are considered when developing and implementing such systems if they are to be accepted and hence effective.

Introduction

Despite the reported benefits of the computer-aided assessment approach, high staff/student ratios often mean that tutors are unable to provide learners with feedback on assessment performance that is timely and meaningful. Freeman & Lewis (1998) amongst others have reported on the importance of feedback as a motivator for student learning. Thus, there is an increasing demand for the development of software applications that would enable the provision of timely, individual and meaningful feedback to those learners who are assessed via computer-aided assessment applications.

In earlier work, we have shown that a system of automated feedback, based on student performance in a Computer Adaptive Test was useful, efficient and generally well regarded by students (Lilley and Barker 2002; 2003; 2004). Barker and colleagues (2002) noted the importance of all major stakeholders in design, implementation and evaluation of projects related to online learning. For this reason, it was important to consider also the views and attitudes of teaching staff to the provision of automated feedback based on a CAT. For this reason, three studies were undertaken to obtain detailed views and suggestions related to our automated feedback prototypes. A summary of the sessions is presented below.

Session 1 Group of 10 computer scientists, teachers, experts in software design and also interested in the provision of online educational systems. A 30 minute presentation followed by a 30 minute moderated, focussed discussion.

Session 2 Group of 50 university lecturers at university conference presentation on MLE. A 25 minute presentation followed by a 5 minute question session and a short questionnaire.

Session 3 Group of 20 university teachers interested in online and blended teaching and learning underwent a 30 minute presentation and 30 minute moderated, focussed discussion.

Each of these sessions each involved a short presentation of the automated feedback prototype, including sample output screens, examples of feedback and also research data related to student performance and attitude to the feedback provided. After each presentation, a semi-structured question and answer session was conducted, where researchers and staff could exchange ideas. Sessions were moderated by an experienced researcher and discussion topics were focussed, based upon a previously prepared script.

The sessions, however, were semi-structured, since open discussion was encouraged on any topic related to the discussion topics.

Sessions were recorded on video and later transcribed in full by the researcher and analysed, using QSR N6 software, in order collate and link together themes and ideas. Responses on the questionnaire administered in session 2 were summarised and is presented below in table 3.

Computer-Adaptive Test (CAT) Prototype Employed in this Study

The development of the CAT application that was the subject of this study has been reported by Lilley and colleagues (Lilley et al. 2004; 2005). The application comprised a graphical user interface, an adaptive algorithm based on the Three-Parameter Logistic Model from Item Response Theory (Lord, 1980; Hambleton, 1991; Wainer, 2000) and a database of questions. This contained information on each question, such as stem, options, key answer and IRT parameters. In this work, subject experts were employed for question calibration. The subject experts used Bloom's taxonomy of cognitive skills (Pritchett, 1999; Anderson & Krathwohl2001) in order to perform the calibration. Questions were first classified according to cognitive skill being assessed. After this initial classification, questions were then ranked according to difficulty within each cognitive level. Table 1 summarises the three levels of cognitive skills covered by the question database and their difficulty range. It can be seen from Table 1 that knowledge was the lowest level of cognitive skill and application was the highest. An important assumption of our work is that each higher level cognitive skill will include all lower level skills. As an example, a question classified as application is assumed to embrace both comprehension and knowledge.

Difficulty *b*	**Cognitive Skill**	**Skill Involved**
$+1 \leq b \leq +3$	Application	Ability to apply taught material to novel situations
$+1 < b < -1$	Comprehension	Ability to interpret and/or translate taught material
$-1 \leq b \leq -3$	Knowledge	Ability to recall previously taught material

Table 1: Level of difficulty of questions

At the end of each assessment session, questions were re-calibrated using response data obtained by all participants who attended the session. In general terms, questions that were answered correctly by many test takers had their difficulty levels lowered and questions that were answered incorrectly by many test takers had their difficulty levels increased.

Our Approach to the Provision of Automated Feedback

It was one of our assumptions that a tutor-led feedback session would typically comprise the provision of an overall score, general comments on

proficiency level per topic and recommendations on which concepts within the subject domain should be revised. It was then planned that the feedback would be made available via a web-based application.

Overall Score

The overall score, or overall proficiency level, would be estimated by the CAT algorithm using the complete set of responses for a given test-taker and the adaptive algorithm introduced in section 2.1. Figure 1 illustrates how this information was displayed within our automated feedback prototype.

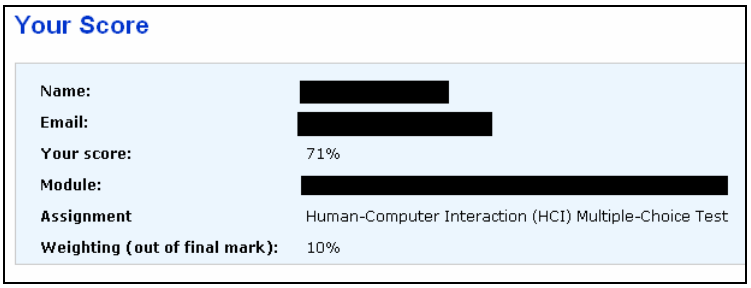


Figure 1: Screenshot illustrating how overall score was displayed within our automated feedback prototype. The student's name and module have been omitted.

Performance Summary Per Topic

Test-takers' responses would be grouped by topic and a proficiency level calculated for each set of topic responses. Proficiency level estimates per topic would then be mapped to Bloom's taxonomy of cognitive skills. The underlying idea was to inform learners about their degree of achievement for each topic domain. Some learners reported that they would also like to compare their test performance with the performance of the whole group. This information was also made available in this section of the feedback, as illustrated in Figure 2.

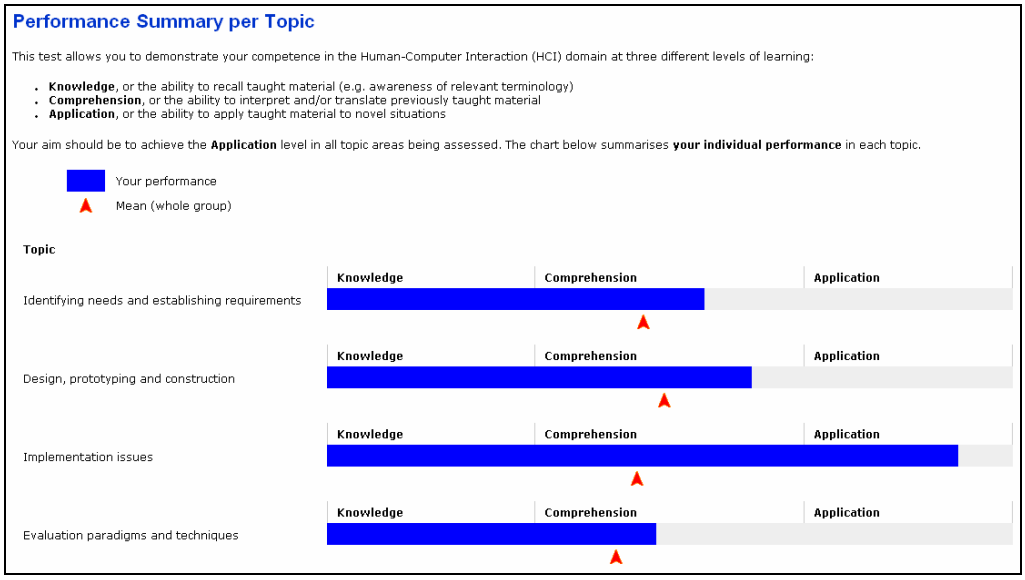
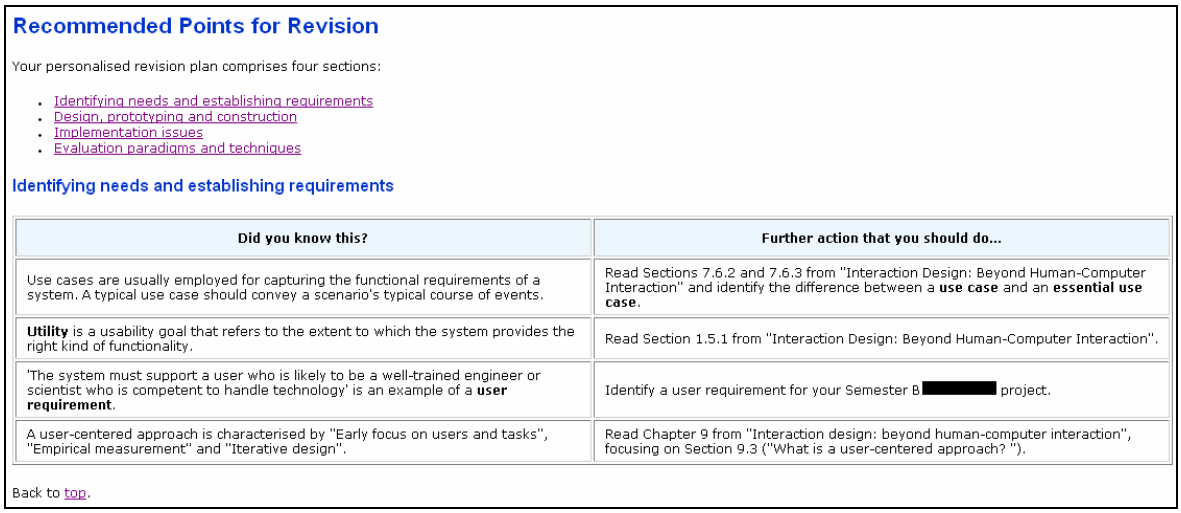


Figure 2: Screenshot of screen containing information regarding performance per topic.

Recommended Points for Revision

An important assumption of our feedback tool was that tutors providing feedback on an objective test during a face-to-face session were likely to provide students with directive feedback rather than simply indicating what the correct options for each question were. As an initial attempt to mimic some aspects of how a subject domain expert would provide learners with recommendations on how to increase their individual proficiency levels, a database of feedback sentences was designed and implemented. This database comprised statements relating to each one of the questions. For each individual student, only those questions answered incorrectly were selected. Figure 3 illustrates the approach to directive feedback employed in this study.



Recommended Points for Revision

Your personalised revision plan comprises four sections:

- Identifying needs and establishing requirements
- Design, prototyping and construction
- Implementation issues
- Evaluation paradigms and techniques

Identifying needs and establishing requirements

Did you know this?	Further action that you should do...
Use cases are usually employed for capturing the functional requirements of a system. A typical use case should convey a scenario's typical course of events.	Read Sections 7.6.2 and 7.6.3 from "Interaction Design: Beyond Human-Computer Interaction" and identify the difference between a **use case** and an **essential use case**.
Utility is a usability goal that refers to the extent to which the system provides the right kind of functionality.	Read Section 1.5.1 from "Interaction Design: Beyond Human-Computer Interaction".
'The system must support a user who is likely to be a well-trained engineer or scientist who is competent to handle technology' is an example of a **user requirement**.	Identify a user requirement for your Semester B [redacted] project.
A user-centered approach is characterised by "Early focus on users and tasks", "Empirical measurement" and "Iterative design".	Read Chapter 9 from "Interaction design: beyond human-computer interaction", focusing on Section 9.3 ("What is a user-centered approach? ").

Back to top.

Figure 3: Example of 'Recommended Points for Revision' for the topic 'Identifying needs and establishing requirements'. The module name has been omitted.

Learner Perspectives on the Usefulness of the Automated Feedback Tool

It was important to ensure that the attitude of learners to the automated feedback tool was positive. In CAA 2005, we provided a report of an evaluation of a feedback session with a group of 113 Computer Science undergraduates participated in a session of summative assessment using our CAT prototype. (Lilley and Barker, 2005). In that study, students received feedback on test performance via the automated feedback tool.

Students then completed a questionnaire in which rated a series of statements using a Likert Scale from 1 (Strongly disagree) to 5 (Strongly agree). A group of 97 students answered the questionnaire and their answers are summarised in Table 2 below.

Question	Strongly disagree 1	2	3	4	Strongly agree 5	Mean	Std Dev
Overall, the feedback tool was effective at providing helpful advice for individual development.	4	5	15	43	30	3.93	1.02
Overall, the feedback tool was effective at providing feedback on performance.	4	4	13	44	32	3.99	1.01
The "Overall Score" section was useful at providing information on how successfully I have learned.	6	9	23	31	28	3.68	1.17
The "Performance Summary per Topic" was useful at providing information on how successfully I have learned in each topic area.	6	6	19	34	32	3.82	1.15
The "Points for Revision" section was useful at providing information on how successfully I have learned.	8	9	14	35	31	3.74	1.24
Overall, I was satisfied with the degree of personalisation offered by the application.	10	7	19	35	26	3.62	1.25
The content of the feedback was appropriate for my individual performance.	6	6	20	39	26	3.75	1.11

Table 2: Learners' perceived usefulness of the feedback approach employed (N=97)

The results shown in Table 3 suggest that the automated feedback approach was favourably received by the learners who participated in the study. It was therefore important to investigate tutors' attitude towards the automated feedback approach proposed here. It was important to be sure that the approach was also acceptable to staff.

Tutors' Perspectives on the Usefulness of the Automated Feedback Tool

Questionnaires

Data obtained in the three sessions reported in section 1.1 was summarised and collated. In the second session, a short questionnaire was administered to provide information on aspects of the automated feedback approach related to formative and summative assessment, objective and essay type tests, and the speed, quality and appropriateness of the approach overall. The answers of 19 tutors who attended the presentation are summarised in Tables 3 and 4 below.

Question	Not useful 1	2	Useful 3	4	Very useful 5	Mean	Std Dev
In the context of summative assessment, the automated feedback approach that I have just seen is:	1	1	10	1	6	3.53	1.17
In the context of formative assessment, the automated feedback approach that I have just seen is:	0	0	8	3	8	4.00	0.94
In the context of objective testing (i.e. multiple-choice questions), the automated feedback approach that I have just seen is:	0	1	7	2	9	4.00	1.05
In the context of written assignments, the automated feedback approach that I have just seen is:	6	5	5	0	3	2.42	1.39

Table 3: Tutors' perceived usefulness of the feedback approach proposed in this study (N=19)

Question	Poor 1	2	Good 3	4	Very good 5	Mean	Std Dev
With regards to its *speed*, the automated feedback approach that I have just seen is:	0	0	4	3	12	4.42	0.84
With regards to its *quality*, the automated feedback approach that I have just seen is:	1	1	8	4	5	3.58	1.12
With regards to its *appropriateness* to enhance students' learning experience, the automated feedback approach that I have just seen is:	1	0	6	4	8	3.95	1.13

Table 4: Tutors' perceived speed, quality and appropriateness of the feedback approach proposed in this study (N=19)

It can be seen from tables 3 and 4 that tutors in general considered the approach to be a useful method for the provision of feedback. This is an important finding, since it will be important that tutors as well as students value the method. Table 3 shows that it is valued more highly in the context of formative, rather than summative, assessment. The use of such automated methods for written assignments was considered the least useful. It was not clear whether this was because of the difficulty of providing automated feedback for written work, or that tutors feel that providing feedback themselves was a better approach. Table 4 shows that on average tutors thought the automated approach to be fast, appropriate and of good quality, though the quality dimension achieved the lowest mean score. All in all tutors' attitude to the approach was positive, which was an important finding.

The Discussion Sessions

In all, three discussion sessions were employed in this study, based on methods described by Barker and Barker (2002). The focus of the second session was primarily to collect the questionnaire data presented in the previous section above. Accordingly there was little opportunity for discussion in this session, which contributed little to the qualitative data obtained. The bulk of the qualitative data obtained in this study therefore was collected in session 1 with a ggroup of 10 computer scientists teachers who were also experts in software design and Session 3 involving a group of 20 experienced

university teachers who were primarily interested in online and blended teaching. In both sessions, after the presentation of our ideas and results, copies of actual feedback (made anonymous) was distributed for inspection. The discussion topics for both sessions are presented in table 5 below.

Discussion topics
What feedback methods do you use at present?
How do you assess the quality of feedback provided at present
What are the limitations and benefits of the feedback you provide currently
What is your view of the CAT approach for formative and summative assessment
What is your opinion of the CAT approach to automated feedback
What are the benefits of the approach
What are the limitations of the approach
How could the automated approach be improved
What should be the role of the tutor in the automated feedback system
What is the need for monitoring and how might this be achieved
What if any are the ethical issues in the method

Table 5: Discussion topics used in focussed sessions 1 and 3

After the presentation on the CAT automated feedback approach by one of the research team, the session moderator introduced the focussed discussion session with a short scripted introduction where the objectives of the discussion and ethical issues, such as confidentiality and the video recording were described to participants. In the first instance, the moderator started the discussion session by asking the first question in table 5, related to the type of feedback provided by tutors at present. Discussion was good in both sessions and for the most part, the moderator merely had to check that all the topics had been covered adequately, and to encourage all present to engage in the discussion where possible. When discussion moved far from the focus, or sufficient time had been spent on a thread, new topics were introduced by the moderator as unobtrusively as possible.

In the first session some discussion by the experts present was related not only to the feedback, but also to the adaptive and modelling ideas related to the software itself. This valuable information was used later, primarily to assist in the software development process in order to improve later iterations of the application. These discussions are not reported in this paper. In the following, a summary of discussions is presented under the topic headings shown in table 5.

Feedback Methods Used at Present

At present, feedback methods employed are mostly classroom and lecture theatre based sessions lasting approximately one hour, given some time after the test, ranging from six weeks to several months. Such sessions are not individual, generally each question is worked through and in some cases, general problems identified by tutors are covered in greater depth. If a question is well answered by most students, then less time is spent on this question. Problem questions are dealt with more fully by most tutors. Other

methods include providing only the questions and worked answers online (either through a web-based system, or by electronic mail). One tutor was using a spreadsheet to attempt to individualise feedback, which amounted to personally typing in comments to the answer sheet for each student. For essay type questions, feedback was usually given as comments written in pen (or sometimes electronically) onto the essay script. Sometimes feedback was provided in small group sessions where topics were discussed, rather than questions analysed in detail. One tutor reported that she used one-to-one sessions to provide feedback on rare occasions. Feedback method seemed to be related to the type of test. For objective tests, most of the methods were employed, with the obvious exception of writing directly on scripts. The purpose of feedback was very much formative, and few reported giving any feedback on summative assessments.

Quality of Feedback Provided at Present

Tutors emphasised the necessity to be able to interact directly with learners and, based upon experience, provide directed and tailored feedback. It was possible to "gauge" how a test had gone, and to provide the necessary feedback in an appropriate format. When pressed as to how this was possible, given large class sizes and the small amount of time devoted to feedback, some tutors agreed that it was not always possible. The quality of feedback provided did indeed vary according to some tutors and inexperienced colleagues might on occasions provide feedback that was variable. When asked to think about the problems of high performing and very low performing students, most tutors agreed that feedback was usually focussed at "the average" student, with an account taken of general problems that appeared in the test itself. Several tutors expressed the opinion that that the quality of our individualised automated feedback was likely to be high, citing the direct feedback on questions answered, the relationship with cognitive aspects of learning as given in the link to Bloom's levels, and the provision of direct online links to more challenging advanced work as well as remedial work based on individual performance on the test. As the feedback was provided in a web format, links to remedial and more challenging materials were active and direct.

Limitations and Benefits of the Feedback Provided Currently

The benefits of the current system might be summarised under the possibility of direct control and monitoring of test performance and feedback. Tutors liked the ability to be able to "keep a finger on the pulse" when providing feedback. Some concern was expressed that an automated approach would lead to potential problems going un-noticed. This could not happen when tutors themselves gave feedback. Some tutors realised that un-timely feedback was far less useful than feedback given quickly. One tutor asked why we imposed a delay in giving out our automated feedback, as it could in theory be sent to students immediately after a test. The need to delay presentation of feedback due to checking and ethical reasons was less likely

to cause undue delay in the future. Most tutors agreed that the speed of the automated feedback was a major benefit.

The CAT Approach for Formative and Summative Assessment

The CAT approach was not the main focus of the discussion, as staff attitude to the CAT aspect had been the subject of earlier studies. It was important however to discuss the CAT in context of the feedback. Most staff were familiar with the CAT approach, as it has been in use in the university for several years now. Benefits of a CAT in terms of efficiency, motivation and plagiarism were already well known. Linking the feedback provided to a CAT was important for us, but not for some other tutors who could see how our automated feedback system could be linked to non-adaptive question banks, though some agreed that there would be a loss of information in such systems, related to the CAT levels in each topic area and the link to Bloom's levels. The use of CAT in summative assessment was generally less well received than for formative testing, which was in accordance with our earlier findings and the questionnaire data from session 2. It was noted by one tutor, however, that the use of a CAT for summative assessment did ensure that timely feedback would be available for all learners at the end of their course, before they had all left the university

The CAT Approach to Automated Feedback

It was realised that the use of automated feedback was an important benefit of the CAT approach. Although some tutors wanted to discuss the CAT approach in greater detail, this was resisted by the moderator and the topic of discussion gently moved. Some tutors expressed the fact that they realised that individual student profiles obtained from a CAT, containing information on performance in topic areas, as well as cognitive levels could be used in a variety of different ways. Some good ideas related to their potential use in teaching and learning were obtained from the session. Some of these are presented in the concluding section of the paper. It was noted that the use of a CAT in automated feedback involves two issues that were closely linked in our study, a CAT and automated feedback. It was our belief, expressed in the presentation, that a CAT was essential to provide individualised and rich automated feedback. It is fair to say that some tutors were not entirely convinced of this link.

Benefits of the Approach and Limitations of Automated Feedback

The most important concern expressed at the sessions related to the loss of control by tutors. Providing automated feedback was liable to remove an important "human aspect" of the teacher's role. The most important benefit cited was the speed of feedback possible with our approach. Other limitations expressed related to the use of objective testing as the only method with the approach and to issues related more to the CAT approach than the feedback itself. Other potential benefits cited included the motivational aspects of CAT

and how this might be used in order to help students do extra work, either remedially, or as extra challenges. This was seen as an important aspect by some tutors. It was emphasised in the presentation prior to discussion that the CAT level obtained represented an important boundary for an individual between what they knew and what they did not know. Providing feedback at this boundary was important and this view was expressed by some tutors present at both sessions. One teacher asked if the profiles obtained in our CAT might be useful in other subject areas. It was possible, due to the objective nature and reproducibility of CAT results that more general information related to learners might be obtained, though we could not confirm this interesting point. Efficiency of the method was also cited as a benefit. Providing feedback in traditional ways was difficult and inefficient as well as being slow. An automated system, once in operation could process test results efficiently with the minimum of human intervention. Admittedly some tutors saw this as a disadvantage, though these were in the minority at both sessions.

Suggested Improvements of the Automated Approach

There were a few suggested improvements to the system. One tutor expressed the opinion that the CAT feedback might be used as the focus either for group seminars or for small remedial classes. It would be possible to obtain useful summaries of strong and week points in the tests in each topic area from the CAT. Such summaries might be useful to tutors in their teaching and for providing remedial materials or lectures. The speed of the CAT would be likely to provide such information quickly and certainly in time for action. Patterns of feedback might be identified in this way and the item database could be analysed to identify problem areas (and areas of strength) in all topics.

The Role of the Tutor in the Automated Feedback System

It is fair to say that a concern of some tutors was that automated feedback was another step on the road to an uncertain impersonal future. This was rarely expressed fully, though it was apparent from some questions that it was a concern. Others expressed the view that there was an opportunity in the approach to develop useful systems that would provide them with more time to develop interesting online and off-computer activities related to the outcome of tests, for example activities related to performance on tests. One teacher suggested that tests could be developed where feedback could be directly incorporated into the CAT and that this might provide a learning opportunity within a CAT. Although outside the scope of our research, this was nonetheless an interesting idea for the future. There would need to be a monitoring role as well as a development role in automated feedback systems and tutors would need to take on this aspect.

Monitoring of the Automated System: Ethical Issues

Our approach to making sure students were not disadvantaged either by our CAT approach or by the way feedback was provided in our system was explained in the introductory presentation. No tutor expressed the feeling that learners would be disadvantaged either by the CAT or by the method of providing feedback as described by us. Most stated the view that it would be important to monitor the CAT and feedback systems to ensure that they were performing properly and fairly. One tutor suggested a method of sampling, both for CAT results and feedback to ensure fairness.

In summary a complex range of issues related to the provision of automated feedback were discussed in these two focussed sessions. Additional information was obtained by means of a questionnaire, completed by attendees at a presentation related to our feedback system. In discussion sessions, tutors were able to explore a range of topics related to how feedback was provided by themselves and colleagues currently and how feedback was provided by our CAT method. In general our approach was well received and tutors were receptive to the ideas in general. They were able to see potential benefits in terms of speed and efficiency and also the ability to personalise feedback at a time when online learning is becoming increasingly important in Higher Education and staff time for providing individual feedback is decreasing. Concerns related to the provision of automated feedback were general in nature, rather than specifically directed at the system we presented. These tended to be focussed on the loss of human input into the system. There was no evidence from these sessions that feedback currently provided by tutors was of a universally high standard or that it was individualised. Rather the contrary opinion was mostly expressed.

Discussion

Substantial investments in computer technology by Higher Education institutions and high staff/student ratios have led to an increased pressure on staff and students to incorporate electronic methods of learning and teaching. This includes a growing interest in the use of computer-aided assessment, not only to make the technological investment worthwhile but also to explore the opportunities presented by the computer technology available. It is our experience that - given the great deal of computerised objective testing that currently takes place – using adaptive tests is an interesting, fair and useful way of providing such assessment (Barker & Lilley, 2003; Lilley et al., 2004). Not only is this motivating for learners, who are challenged appropriately - i.e. not discouraged by questions that are too hard, or de-motivated by questions that are too easy - but also the information that it provides can be used in interesting and useful ways. For instance, it can be used in the presentation of remedial work for students or, as in our case, for the provision of personalised feedback.

Feedback must be timely to be useful. Our experience is that when large-scale computerised objective testing is used in a formative context, results are usually returned quickly, because of automated methods of marking. Feedback, however, is often slow and delivered by the time the course has moved on and it is of less use or, in some cases, feedback is absent. This experience was largely confirmed by the results obtained in the current study. It is time consuming to produce individual feedback for hundreds of students. When feedback is provided, it is usually little more than a final score, generic worked examples and a list of questions answered correctly and incorrectly. Automated methods are therefore likely to be useful in this context, as evidenced by the tutors' attitude reported in this study. The matching of adaptive testing and automated feedback provides an opportunity to individualise feedback to a far greater extent. We argue that the automated feedback approach proposed here, which is based on adaptive testing, is appropriate for identifying learners' strengths and weaknesses for each topic area covered by the test. Automated feedback as proposed in this study is also related to Bloom's levels, thus providing meta level information for learners about the depth of their approach in each of the topic areas. This information would be difficult to obtain with standard objective testing.

Other approaches to the provision of feedback to groups of learners, such as in-class sessions where all questions from an objective test are presented by a tutor, are likely to remain as important feedback methods. Such in-class approaches offer high quality information about the test and each of the questions, often providing learners with an opportunity to work through the questions. They do not, however, address the individual needs of many of the learners. Explaining a question that is set at a difficulty level that is too low for most learners will not be of interest for the majority of the group. Similarly, it can be argued that discussing questions that only one or two learners are capable of answering will not be the most efficient way of employing tutors' and learners' time. We suggest that not only is the automated feedback based on adaptive testing a fast and appropriate method, but that it also provides information to learners that would be difficult to obtain elsewhere, given the decrease in the number of face-to-face sessions, the increase in staff/student ratios and the growing trend in the use of electronic resources for the delivery of courses, assessment, student feedback and support.

Our research has shown that learners and tutors accept and value the automated feedback approach proposed in this study. In the future we intend to apply this method more widely, for example in providing feedback for written assignments. We also intend to use the wealth of information about learners' proficiency levels provided by the adaptive testing approach to develop useful student models. Such student models will, in turn, be employed to generate profiles that could be used in a wide variety of learning contexts.

References

Anderson, L. W. & Krathwohl, D. R. (Eds.) (2001) *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, New York.

Barker, T. & Barker, J. (2002) "The evaluation of complex, intelligent, interactive, individualised human-computer interfaces: What do we mean by reliability and validity?", *Proceedings of the European Learning Styles Information Network Conference*, University of Ghent, June 2002.

Freeman, R. & Lewis, R. (1998) *Planning and Implementing Assessment*, Kogan Page, London.

Hambleton, R. K. (1991) *Fundamentals of Item Response Theory*, Sage Publications Inc, California.

Lilley, M. & Barker, T. (2002) "The Development and Evaluation of a Computer-Adaptive Testing Application for English Language", *Proceedings of the 6th Computer-Assisted Assessment Conference*, Loughborough University, United Kingdom, pp. 169-184.

Lilley, M. & Barker, T. (2003) "Comparison between Computer-Adaptive Testing and other assessment methods: An empirical study", *Proceedings of the 10th International Conference of the Association for Learning Technology (ALT-C)*, University of Sheffield, United Kingdom.

Lilley, M. & Barker, T. (2004). "A Computer-Adaptive Test that facilitates the modification of previously entered responses: An empirical study", Proceedings of the 2004 Intelligent Tutoring Systems Conference, *Lecture Notes in Computer Science 3220*, pp. 22-33.

Lilley, M., Barker, T. & Britton, C. (2004) "The development and evaluation of a software prototype for computer adaptive testing", *Computers & Education Journal 43(1-2)*, pp. 109-123.

Lilley, M., Barker, T. & Britton, C. (2005) "The generation of automated learner feedback based on individual proficiency levels", Proceedings of the 18th International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems, *Lecture Notes in Artificial Intelligence 3533*, pp. 842-844.

Lord, F. M. (1980) *Applications of Item Response Theory to practical testing problems*. Lawrence Erlbaum Associates, New Jersey.

Pritchett, N. (1999) "Effective question design" In S. Brown, P. Race & J. Bull (Eds.), *Computer-Assisted Assessment in Higher Education*, Kogan Page, London.

Wainer, H. (2000) *Computerized Adaptive Testing (A Primer)*, Lawrence Erlbaum Associates, New Jersey.

ISSUES WITH SETTING ONLINE OBJECTIVE MATHEMATICS QUESTIONS AND TESTING THEIR EFFICACY

Nabamallika Baruah, Mundeep Gill and Martin Greenhow

Issues with Setting Online Objective Mathematics Questions and Testing their Efficacy

Nabamallika Baruah, Mundeep Gill and Martin Greenhow
Department of Mathematical Sciences
Brunel University
nabamallika.baruah@brunel.ac.uk

Abstract

The Mathletics database now comprises many mathematical topics from GCSE to level 2 undergraduate. The aim of this short paper is to document, explore and provide some solutions to the pedagogic issues we are facing whilst setting online objective questions across this range. Technical issues are described in the companion paper by Ellis, Greenhow and Hatt (2006). That paper refers to "*question styles to stress that we author according to the pedagogic and algebraic structure of the content of a question; random parameters are chosen at runtime ... This results in each style having thousands, or even millions, of realisations seen by the users.*" With this emphasis, and with new topics being included, new question types beyond the usual multi-choice (MC) etc have been developed to ask appropriate and challenging questions. We feel that their pedagogic structure (and underlying code) is widely applicable to testing beyond the scope of Mathematics. This paper describes three of the new question types: Word Input, Responsive Numerical Input and 4/True/False/Undecidable/Statement/Property. Of generic importance is the fact that each of these question types can include post-processing of submitted answers; sample Javascript coding that checks the validity of the input(s) before marking takes place is described. In common with most of the rest of the question style's content this could be exported to other CAA systems.

Ellis et al (2005) and Gill & Greenhow (2006) describe initial results of a trial of level 1 undergraduate mechanics questions. This academic year we have expanded the range of tests to foundation and level 1 undergraduate algebra and calculus, involving several hundred students. First and foremost we have underlined the value of Random Numerical Input (RNI) question types compared with traditional Numerical Input (NI) types for which answer files resulting from questions with randomised parameters are exceptionally difficult to interpret. Despite our current lack of a consistent and fully-meaningful way of encoding the mal-rules within the question outcome metadata, mal-rule-based question types (MC, RNI etc) have been analysed in terms of difficulty, discrimination and item analysis. In the case of multiple-choice questions any weaknesses are separately identified as skill-based or conceptual.

Introduction

Multiple-choice questions are the most common types of questions used to set objective tests. Previous papers (Gill & Greenhow, 2006; Ellis *et al* 2005; Gill & Greenhow, 2005;) have discussed the methodology we have used at Brunel University to ensure that the options made available in multiple-choice questions are reliable and realistic. Past exam scripts in the areas of calculus and mechanics have been analysed to identify common mistakes that students make while answering certain types of questions. Similar work is also currently being carried out in the area of algebra. It is hoped that by identifying common mistakes and using these as distracters, the feedback will be more focused on individual errors and feedback to the lecturers will also highlight common mistakes that students are making.

Many objective tests have been set up and used at Brunel University over the past academic year. These tests cover areas such as algebra, calculus, mechanics and statistics, mainly at level 0 and level 1. Some tests have been used purely for formative reasons while others have been used for summative purposes. Students are encouraged to use the questions for revision purposes to aid them in their learning process. From analysis of student answer files for calculus and mechanics it was found that a higher percentage of students were able to answer multiple-choice type questions correctly compared with numerical input (see table 2 below). Since final examinations do not generally contain multiple-choice questions, it was decided to develop other types of questions.

Some New Question Types

Word Input (WI)

Even in a tightly-specified setting requiring the input of only short phrases, marking algorithms in any objective system will find it difficult to equate the meaning of equivalent forms (e.g. *x is at least as large as y is equivalent to x is not smaller than y*). We have sought to facilitate the communication between user and marking scheme by casting questions in terms of the positions taken by protagonists. A very simple example is shown in figure 1, but this type could be used to require students to evaluate each of the protagonist's positions on a more complex or incompletely-specified "real-world" problem. Figure 1 shows a situation with five possible answers (note the use of *Nobody*), since here we need to link names with a mathematical expression; we have effectively created a multiple-choice question in another form. However, it would be entirely feasible to set up a much less constrained question stem with an arbitrary number of (unique) names, asking, for example, who's position is best supported by the evidence presented.

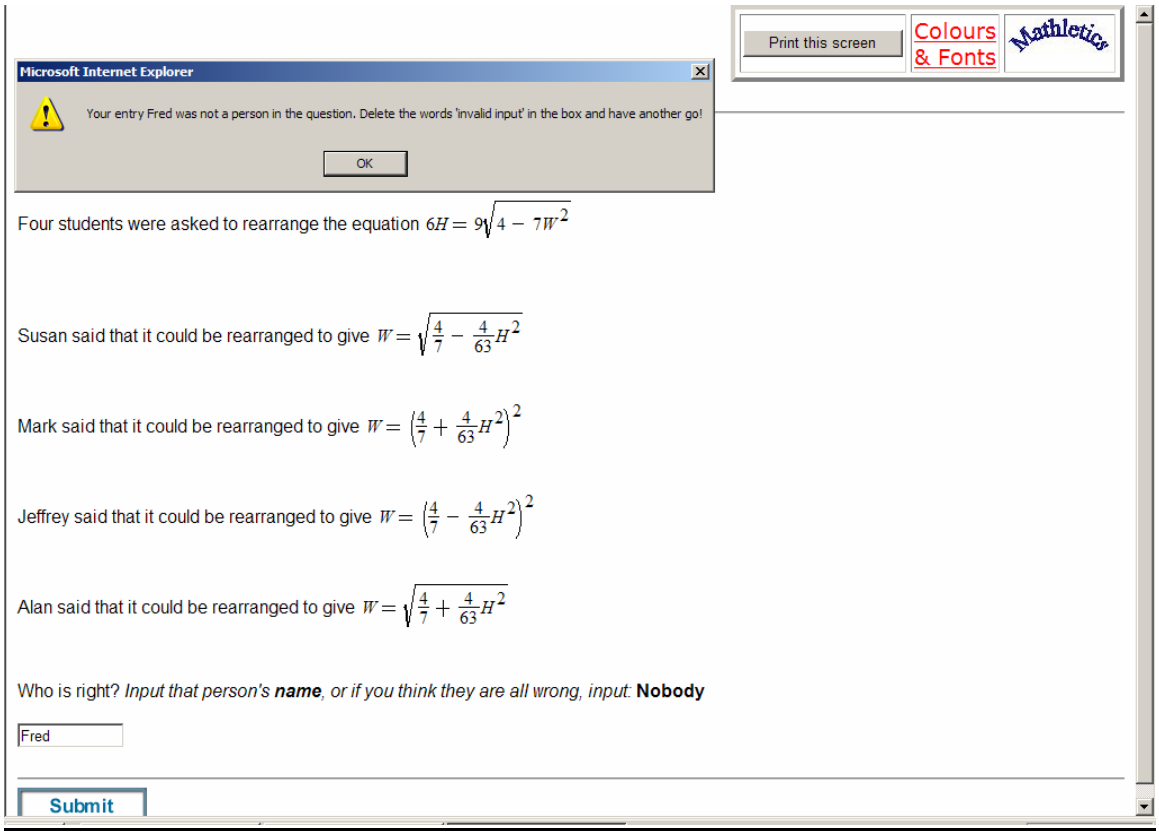


Figure 1 A

The variable names (H and W) are randomly chosen from a subset of upper/lower case alphabetical characters. All numbers are randomised with certain bounds determined by the pedagogy of the question (e.g. how difficult should the arithmetic be?). The protagonists' names are selected randomly from male/female datasets reflecting the 16-25 year old UK ethnic mix. This results in millions of (pedagogically and algebraically equivalent) realisations of this question style.

Although seemingly straightforward to mark, a degree of post-processing of user input is now required. By comparison with each of the n entries in the question's protagonist list (person[]), we firstly check that an entry is a valid name (not a misspelling or in the wrong case) and issue an appropriate warning (as shown in figure 1) if necessary. Next, for the sake of correct grammar, proper nouns are automatically capitalised for the student if they have not used them before marking comparison takes place. We believe that something like the following code will generally be needed for robust handling of word input:

```
//If input did not begin with an upper case, then this will be automatically updated for them
okinput=0
for (k = 0; k <=n-1; k++){
if (document.forms[0].elements[item].value.toUpperCase() == person[k].toUpperCase()){okinput=okinput+1}
}

if (document.forms[0].elements[item].value.toUpperCase() == "nobody".toUpperCase()){okinput=okinput+1};

//If input was not a person in the question, then a alert message is prompted saying so
if (okinput == 0) {
        alert("Your entry "+document.forms[0].elements[item].value+" was not a person in the question. Delete the
words 'invalid input' in the box and have another go!");
document.forms[0].elements[item].value="invalid input";}else{

strlength = document.forms[0].elements[item].value.length;
part =  document.forms[0].elements[item].value.substring(0,1).toUpperCase();
rest =  document.forms[0].elements[item].value.substring(1,strlength).toLowerCase();

document.forms[0].elements[item].value = part+rest}
```

Responsive Numerical Input (RNI)

A weakness of basic numerical input type questions is that the answer inputted by students is marked either correct or incorrect. Therefore the feedback provided can only indicate whether students answered the question correctly or not and provide the standard worked solution. These types of questions do not provide directed feedback, as multiple-choice do, and hence are not seen to be as effective. However, we have developed a new question type known as *Responsive Numerical Input*. This type of question is very similar to multiple-choice but differs in that (an arbitrary number of) distracters are coded in the background and are not presented to students as in multiple-choice questions. This means that if a student makes a particular mistake that has been coded as a mal-rule, then the feedback can be similar to that of a multiple-choice question, correcting specifically the mistake they have made in their working; for example, a student may have interpreted (a+b)/c as a+b/c. Partial credit can be awarded if appropriate. However, in contrast to multiple-choice questions, students will be unable to eliminate the correct answer from a list of options. Feedback to lecturers will be more informative and students will be faced with a more realistic form of testing, i.e. similar to that of exams.

Responsive numerical input type questions can also be extended to *Sequential Responsive Numerical Input* types. This type of question is used for questions that contain more than one part and the different parts are connected. For example, students may need their answer to the first part to answer the second part. The advantage of using a sequential responsive numerical input type is that not only will feedback be directed (as in responsive numerical input) but students can also be told whether the method they attempted is correct or not (given their answer to the previous part of the question was incorrect). Figure 2 shows an example of a sequential responsive numerical input type question.

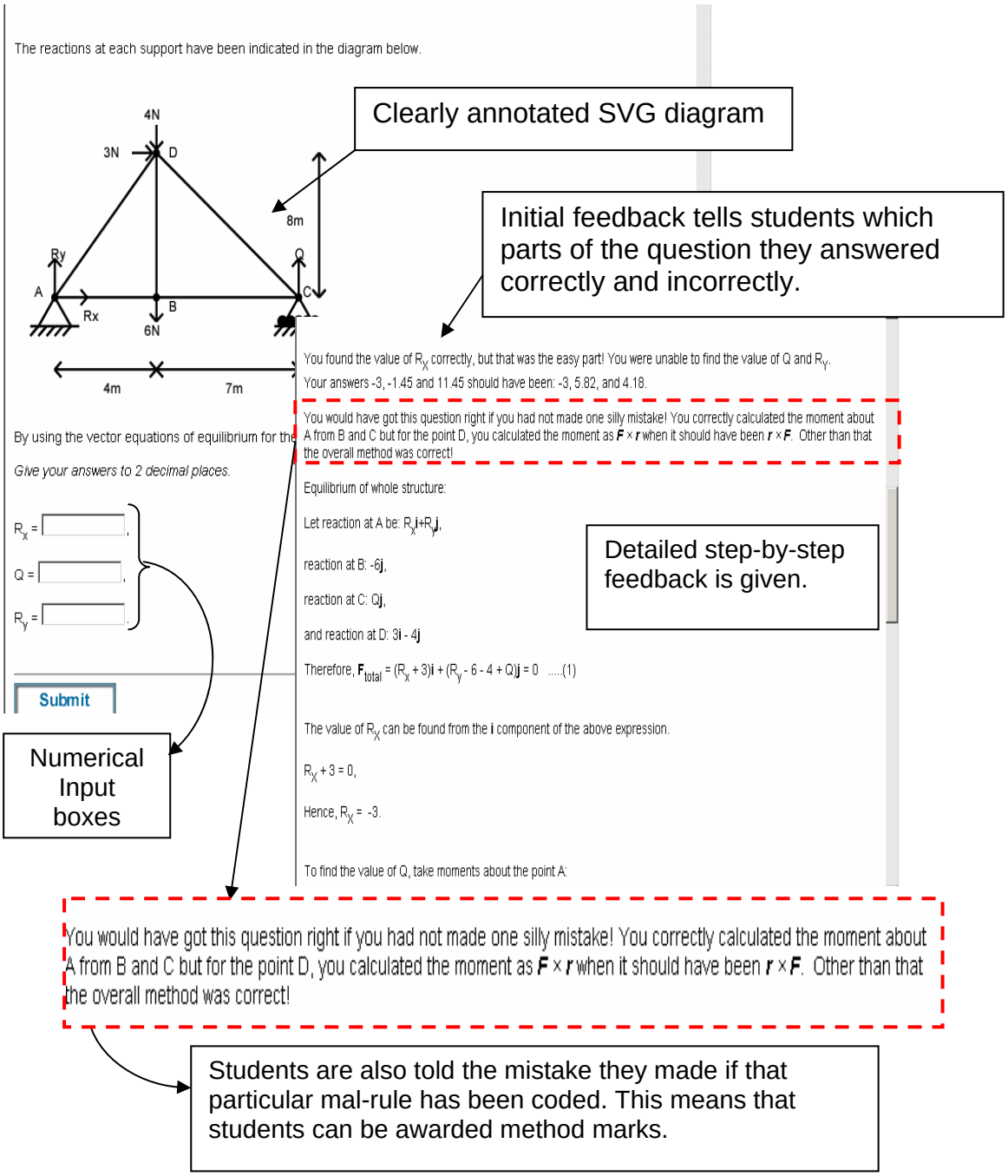
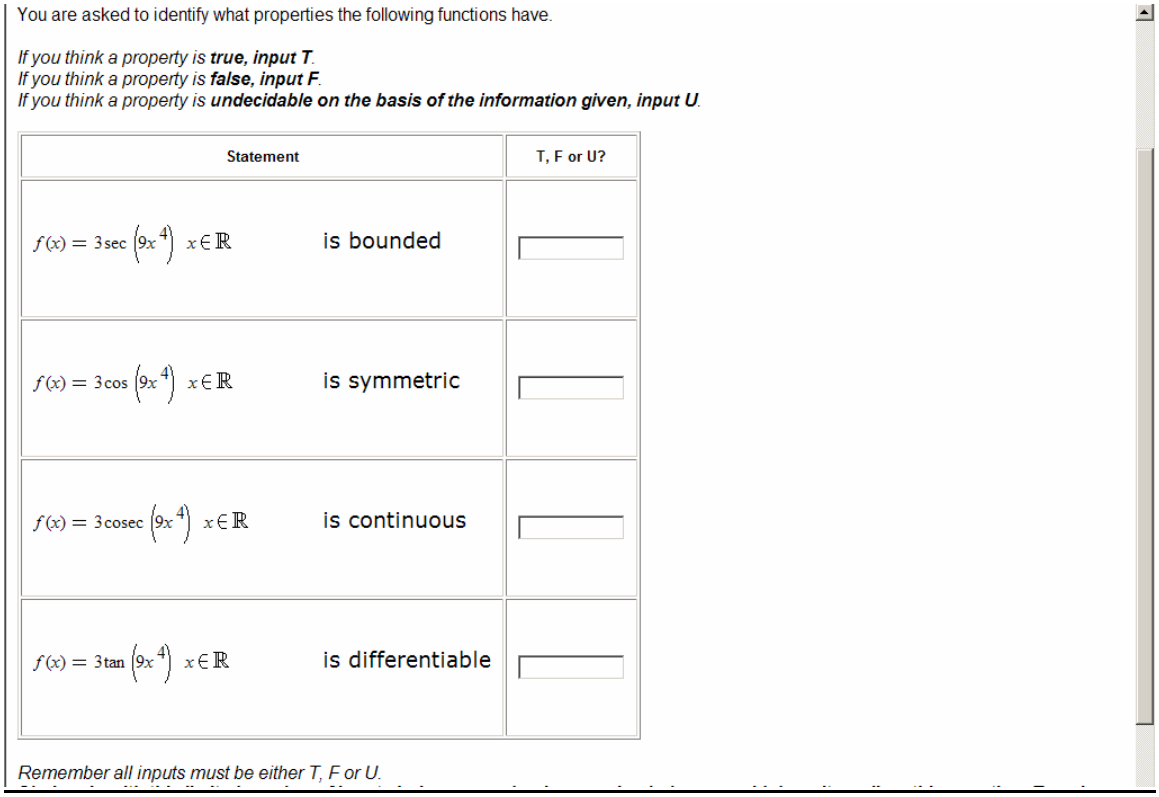


Figure 2: Example of a Sequential Responsive Numerical Input type question

The feedback that is provided to students not only indicates the parts of the questions that students answered correctly and incorrectly, but it also tells students where an error in their working has been made. This type of feedback is useful for questions where the method students are required to use is lengthy and students may spend a long time attempting such questions. The amount of coding required for a question such as that shown in Figure 2 is extensive, but it is hoped that students find such questions worthwhile and more challenging than multiple-choice type questions.

4 True, False or Undecidable; Statement and Property (4TFUSP)

Figure 3 shows a realisation of this type of question. Not only are the statement parameters (choice of trig function and coefficients) randomised, but the properties of the propositions (bounded, symmetric etc) are also randomised. This considerably expands the number of realisations available in the question style. By adding four parts to the question an expansive almost exam-like question is generated that could challenge many students. Variants having either statement of property choice fixed, are useful for determining a students' knowledge of a function (e.g. sine having properties such as continuity, antisymmetry etc) or a property (e.g. which of the randomly-chosen functions are symmetric).


You are asked to identify what properties the following functions have.

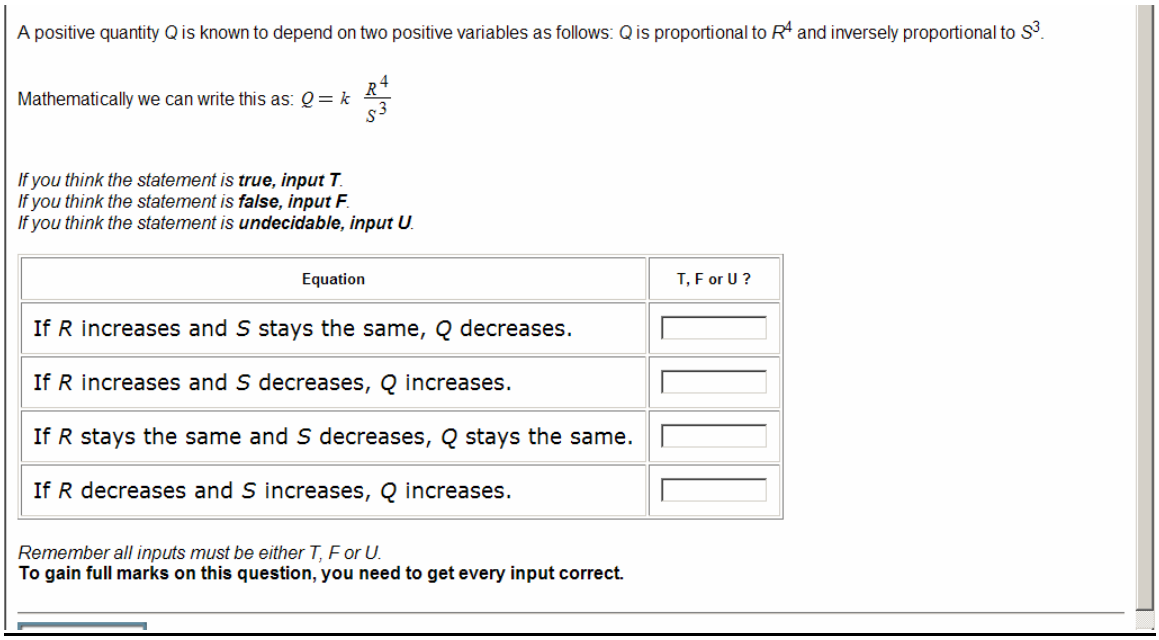
*If you think a property is **true, input T**.*
*If you think a property is **false, input F**.*
*If you think a property is **undecidable on the basis of the information given, input U**.*

Statement		T, F or U?
$f(x) = 3\sec\left(9x^4\right)\ x \in \mathbb{R}$	is bounded	
$f(x) = 3\cos\left(9x^4\right)\ x \in \mathbb{R}$	is symmetric	
$f(x) = 3\operatorname{cosec}\left(9x^4\right)\ x \in \mathbb{R}$	is continuous	
$f(x) = 3\tan\left(9x^4\right)\ x \in \mathbb{R}$	is differentiable	

Remember all inputs must be either T, F or U.

Figure 3. A 4 True, False or Undecidable; Statement and Property (4TFUSP) question type.

Another example is shown in figure 4. Obviously the question stem could be altered to describe a "real-world" scenario with the input boxes stating plausible conclusions or recommendations that might, or might not, follow from the scenario. Indeed it is planned to utilise this type of question (and word input questions) to test students' understanding of statistical inference and transferable skills, such as critical thinking. Notice again that the validity of student input must be checked, with lower case t, f, u inputs being changed to capitals. All other inputs triggering an invalid input message similar to that shown in figure 1 must be addressed.



A positive quantity Q is known to depend on two positive variables as follows: Q is proportional to R^4 and inversely proportional to S^3.

Mathematically we can write this as: $Q = k \frac{R^4}{S^3}$

*If you think the statement is **true, input T**.*
*If you think the statement is **false, input F**.*
*If you think the statement is **undecidable, input U**.*

Equation	T, F or U ?
If R increases and S stays the same, Q decreases.	
If R increases and S decreases, Q increases.	
If R stays the same and S decreases, Q stays the same.	
If R decreases and S increases, Q increases.	

Remember all inputs must be either T, F or U.
To gain full marks on this question, you need to get every input correct.

Figure 4. A 4 True, False or Undecidable; Statement and Property (4TFUSP) question type testing interpretation of a mathematical expression.

Methods Used to Evaluate the Feedback Provided and the Overall Question Efficacy

For all questions that have been produced much time and effort has been dedicated to the feedback being provided to the students. Within the Brunel group there was much debate over the amount of feedback that should be provided: some members thought that students would simply ignore the feedback if too much was provided, while others thought that students would benefit from the detailed feedback. We therefore decided to investigate how effective the feedback provided actually was. Initial results, mainly specific to the topic area of mechanics, were reported in Gill & Greenhow (2006); we now have more data to report.

Over the past two academic years we have incorporated mechanics lab sessions into the level 1 mechanics module at Brunel University (a core module for Mathematics students). These sessions ran on a weekly basis and though not compulsory, the students were encouraged to attend. Students completed a different assessment at each session, and were able to make use of any resources they wanted. Answer files for all assessments attempted were also recorded. We used the Assessment Experience Questionnaire (AEQ), from the Formative Assessment in Science Teaching (FAST) project group (FAST 2004), to get very positive feedback from the students about the questions, see Gill and Greenhow (2006). That paper also identifies the longer-term effects of participation in the lab sessions on students' approach to tackling questions on the end-of-module exam.

Student Retention Periods: Recorded Answer Files

It was hoped that although the feedback provided was extensive, students would be able to retain and make use of it after a delayed time period. Some students repeated the assessments more than once, either within the same lab session or after a period of time. By analysing these student answer files we aimed to see if students could retain the feedback and make use of it in their subsequent attempts. Table 1 shows the results obtained from the analysis of student answer files for mechanics topics: no similar data is yet available for calculus or algebra topics. It lists each assessment that students repeated and the periods of time students were able to retain the feedback. These have been grouped into either short time periods (1 day to 4 weeks) or long time periods (5 weeks to 7 weeks).

Assessment	Retention Period			
	Retain Feedback Immediately	Short Period 1 day to 4 weeks	Long Period 5 weeks to 7 weeks	Unable to retain feedback for any period of time longer than immediate use
Forces & Vectors	6	1	2	5
Forces & Vectors 1	5	1	2	3
Resolving Forces	3		1	2
Resolving Forces (Tension)	3	4		6
Resolving Forces (Equilibrium)	4	1	3	2
Resolving Forces (Inclined Plane)	5	1	1	4
Revision of Resolving Forces	2			1
Trusses & Loaded Beams	3	1		4
Trusses	2			4
TOTAL	**33**	**9**	**9**	**31**

Table 1: Retention of feedback as identified by correct answers recorded for subsequent test(s) for each of the topic areas; from Gill and Greenhow (2006).

On analysing student answer files it was found that all students were able to retain the feedback long enough to make use of it within the same day. However, many students were unable to retain the feedback for any longer other than immediate use. Some students were able to retain the feedback for a period of 7 weeks, which may imply that these students have mastered the material that was being tested. These results are positive and imply that students are able to retain the feedback provided to them. Observations made

during the lab sessions indicated that many students were using the questions as a learning tool rather than an assessment. There was evidence of randomly selecting options and inputting random numbers just to get to the feedback screen. This was surprising since it was thought that students would be more concerned with what *mark* they received and would therefore make use of other resources to help them answer the questions. In actual fact students made use of the questions by reading through the feedback and then reattempting them.

Item Analysis

Mechanics assessments

Throughout all the mechanics assessments there were 2 main question types: Multiple-choice and Numerical Input. The numerical input questions ranged from 1 numerical input to 4. Some questions were sequential and/or responsive. So far we have only analysed the results in terms of students answering the different types correctly and incorrectly. Individual question item analysis has yet to be done where common student mistakes can be identified and reported on. Table 2 shows the percentage of students that answered the different question types correctly and incorrectly.

Question Type	Correct	Distracters	Other (Don't know or only parts correct)	Wrong	Random Input for Feedback
Multiple-Choice	58%	21%	9%	12%	-
1 Numerical Input	38%	-	-	62%	-
2 Numerical Input	39%	-	18%	43%	-
3 Numerical Input	20%	4%	35%	24%	17%
4 Numerical Input	3%	11%	11%	50%	25%

Table 2: Summary of ways students answer different question types.

Table 2 shows that a higher percentage of students answer multiple-choice questions correctly compared with the other types of questions. One possible reason for this may be due to the fact that 4 numerical options are presented to select the answer from (although n*one of these* could be the correct answer). Students have the opportunity to work through a number of different methods until they have a numerical answer that is identical or at least similar to one that is presented to them. In a sense this makes multiple-choice questions 'easier' to attempt compared with Numerical Input types and hence strengthens the need to use question types such as Responsive Numerical Input.

Roughly the same percentage of students answer 1 Numerical Input and 2 Numerical Input types correctly. Many students did not even attempt to

answer 3 Numerical and 4 Numerical input type questions but used them only for the purpose of reading through the feedback.

Foundation level assessments

The item facility index is one of the most useful, and most frequently reported, item analysis statistics. The facility index of an item indicates what percentage of students know the answer. For this reason it is frequently called the *p-value*.

Table 3 shows a small selection of questions that were used to test 170 foundation students on differentiation and integration. The table indicates the concept being tested, facility of the question and the discrimination.

Question Type	**Concept**	**Facility**	**Discrimination**
Multiple-Choice	Differentiation: Chain rule	0.629	0.815
	Differentiation: Product rule	0.551	0.554
	Integration: Polynomial	0.71	0.669
	Differentiation: Polynomial	0.667	0.702
RNI	Integration: Rational form	0.363	0.447
	Integration: Polynomial form	0.34	0.753
	Integration: Powers	0.273	0.805
NI	Integration: Logarithmic form	0.056	0.472
	Differentiation chain rule	0.417	0.789
Hot line	Differentiation chain rule	0.407	0.615

Table 3: A selection of questions that were used in the foundation differentiation and integration test.

The facility of the multiple choice questions range from 0.551 to 0.71. This indicates that students did not find these particular questions difficult or challenging. In comparison, students found responsive numerical input questions difficult since the facility ranged from 0.273 to 0.363. This is much lower than the facilities obtained for the multiple choice questions. Similarly, numerical input questions were also perceived to be difficult since the facility levels ranged from 0.056 to 0.417. This indicates that numerical input type and responsive numerical input types are comparatively harder than multiple choice questions.

Discrimination measures how performance on an item correlates to performance in the test as a whole. There should always be some correlation between item and test performance, however, it is expected that discrimination will fall in a range between 0.5 and 1.0. Figure 5 shows the relationship between discrimination and facility for the results obtained from the integration test.

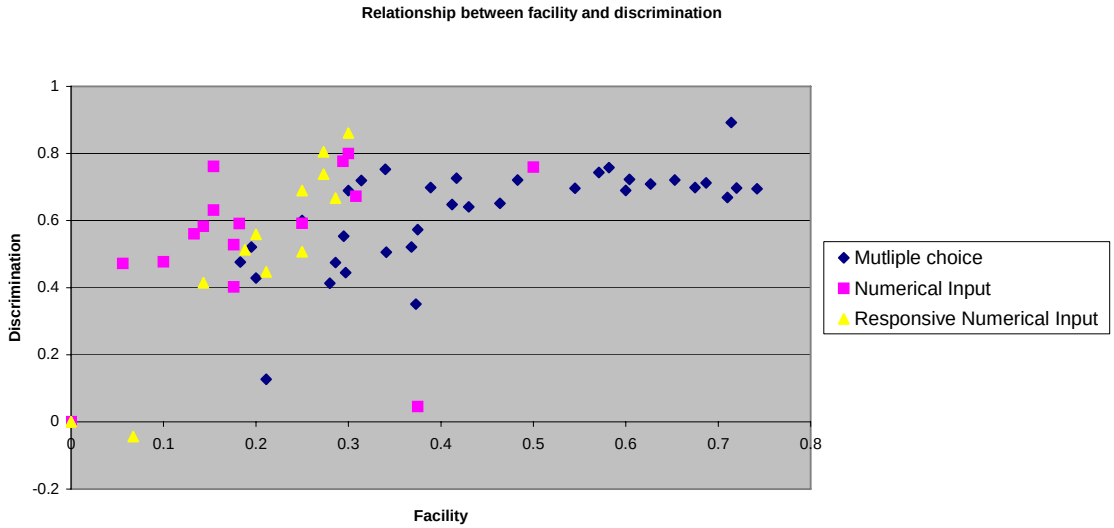


Figure 5: A scatter diagram of the relationship between facility and discrimination for questions in the foundation integration test

From Figure 5 it can be seen that differing facilities between the question types is apparent. The facility for numerical and responsive numerical input type questions is small whereas the mean for the multiple choice questions is much larger. For the majority of the questions the discrimination level is above 0.4, which indicates that most of the questions discriminate well, and ensured the efficacy of the test. The items lying above discrimination level of 0.5 indicate that these questions are highly discriminating.

The items showing negative correlation indicates that a higher proportion of the low scoring group answered the question correctly than that from the high scoring group and conversely. Such type of questions should be examined for finding the possible reason(s) for the reverse difference between the high and low scoring groups.

In the case of multiple-choice questions, responsive numeric input and hot line questions the weaknesses can be separately identified as skill-based or concept based. The structured mal rules record the difficulties of the students in the answer file. Before setting the questions, their objectives are determined (whether skill based or concept based). The skill level and the concept level questions of the foundation level calculus test has been analysed according to the mean facility and the discrimination index.

Levels	Mean facility	Mean discrimination index
Skill	0.48	0.48
Concept	0.475	0.467

Table 4: Table showing mean facility and mean discrimination index for skill and concept questions.

It has been observed that the mean facility and discrimination of the two levels i.e. skill and concept are nearly equal. The lower difference of facility and discrimination of both the skill based and concept based question indicate that the questions are of moderate difficulties with acceptable discrimination.

Conclusions

Our results so far show considerable variability of success rate for different question types across a range of mathematical topics. Students certainly engage with the questions and make extensive use of the feedback provided; they regard this as a valuable learning resource and appreciate the directed feedback offered in response to wrong choices made for multiple-choice questions. Therefore, as part of a formative assessment, multiple-choice questions are very valuable in building knowledge and confidence. However, comparison with other question types, such as numerical input, show the limitations of multiple-choice questions when used summatively or for testing topic mastery. This implies that a variety of question types, including the new ones described here, should be used to give a more sophisticated measure of the student's profile of skills and abilities. In particular we recommend that responsive numerical input types should displace traditional numerical input questions, and multi-stage questions should be authored as sequential (responsive) numerical input if possible.

References

Ellis, E., Baruah, N., Gill, M., Greenhow, M. 2005 Recent developments in setting objective tests in mathematics using QM Perception Proc 9th CAA Conference, Loughborough, July http://www.caaconference.com

E Ellis, M Greenhow, Hatt, J. 2006 Exportable technologies: MathML and SVG objects for CAA and web content Proc 10th CAA Conf, Loughborough, July. http://www.caaconference.com/

FAST – Formative Assessment in Science Teaching 2005 http://www.open.ac.uk/science/fdtl

Gill, M. & Greenhow, M. 2004, Setting objective tests in mathematics using QM Perception Proc 8th CAA Conference, Loughborough, July http://www.caaconference.com

Gill, M. & Greenhow, M. 2005, Learning via online mechanics tests Proc Science Learning and Teaching Conference, Warwick, June

Gill, M. & Greenhow, M. 2006, Computer-Aided Assessment in Mechanics: what can we do; what can we learn; how far can we go? Proc IMA Conf Mathematical Education of Engineers, Loughborough, April.

A DIAGRAM DRAWING TOOL FOR SEMI–AUTOMATIC ASSESSMENT OF CONCEPTUAL DATABASE DIAGRAMS

F.Batmaz and C.J.Hinde

A Diagram Drawing Tool for Semi–Automatic Assessment of Conceptual Database Diagrams

F.Batmaz and C.J.Hinde
Research School of Informatics
Computer Science
Holywell Park
Loughborough University
F.Batmaz@lboro.ac.uk
C.J.Hinde@lboro.ac.uk

Abstract

The increased number of diagram based questions in higher education has recently attracted researchers to look into marking diagrams automatically. Student diagrammatic solutions are naturally very dissimilar to each others. However, it has been observed that there are a number of identical diagram components. This observation forms the basis of our semi–automatic assessment. Identifying identical diagram components in student diagrams needs contextual information about each component. This paper proposes a diagram tool which obtains the contextual information of each component in a conceptual database diagram.

Introduction

Automatic marking of student conceptual database diagrams is a difficult problem like free text marking [1]. However, the assessment process can be altered to make it suitable for automation as long as that alteration is justified educationally. This research investigates requirements of the assessment environment, which can help the examiner during the marking by analysing the existing manual assessment in order to computerise it as much as possible. It is believed that this approach will form the foundation for fully automated assessment. In addition, the research results have some immediate practical uses.

This research focuses on semi-automatic diagram marking. The aim of semi-automation is to reduce the number of sub-diagrams marked by the examiner. This requires identifying and grouping identical sub–graphs in student solutions. This is a similar approach to the Assess by Computer (ABC) Project [2], however the approach used for grouping the diagrams in our research is very different from the ABC Project. The ABC project defines identical components by using those component's attributes (e.g. label, type, Adjacent Boxes). In our research, identical components are defined by the references

to the text describing the scenario. A similar approach is used for intelligent tutoring system in the KERMIT project [3].

The ABC and KERMIT projects have developed their own diagram editors to capture student diagrams. This research also requires its own diagram editor, which is discussed in the diagramming tool section. A prototype of the diagram editor has been tested on students. Results from this may be found in the experiment sections and further work is described in the final section.

Related Work

There are four other recent studies known [1,2,3,5], which are concerned with automatic assessment of conceptual database diagrams. However, there have been many other studies on automatic production and integration of conceptual diagrams. These could be directed at automatic assessment, but are not addressed here.

The DEAP Project [1] at The Open University uses statistical techniques to grade student exam scripts. This work likens imprecise diagrams to free-form text. The associated commercial intelligent free-form text assessor uses latent semantic analysis for marking [4]. In this analysis, to perform a semantic matching between student text and ideal solution, the semantic of a word is determined from the paragraph in which that word occurs. The DEAP Project looks for suitable keywords in student answers to mark free-form text. It has considered a “relationship" in E-R diagram equivalent to a word in text and applied the same statistical technique to grade the diagrams. Their initial results show that the automatic grading of simple diagrams is feasible.

The ABC Project [2] aims to present student design to the human marker after filtering out diagrams which are identical so that the speed and quality of the marking process can be improved. ABC uses graph isomorphism with some heuristics for local metrics of matching diagrams. It is reported that the approach works well on large, artificial, examples, but tests with real examination data produced some unexpected results. The results have shown some matches which are not actually valid (over-match). In their approach, matching is largely dependant on the component labels.

DATsys [5] is part of the Ceilidh system and provides a customizable environment to create various kind of diagrams. Model answers and student diagrams are captured by DATsys and then another Ceilidh module marks the diagrams. The Ceilidh system was originally designed for assessing programming. The system marks, for instance, a student flowchart diagram by first converting the diagram into a BASIC program and then checks the program against the test data. DATsys hasn’t been used to assess ER Diagrams yet. There is some very early stage research of adapting DATsys for ER diagram marking [6].

KERMIT [3] is an intelligent tutoring system aimed at the university-level students learning conceptual database design. KERMIT contains a set of problems and ideal solutions to them. Unlike traditional ITS, it hasn't got a problem solver. The system compares the student solutions to the ideal solution using domain knowledge represented in the form of constraints, which are classified into syntactic and semantic ones. The semantic constraints enable the system to deal with alternative student correct solutions. Correspondences between the components of the student and the ideal solution are found by forcing the student to highlight the word or phrase in the text whenever a new part is added to the diagram. These correspondences are used to fire the appropriate production rule/s in the semantic constraints. In the case of violation of any of these constraints, feedback is generated.

Approach

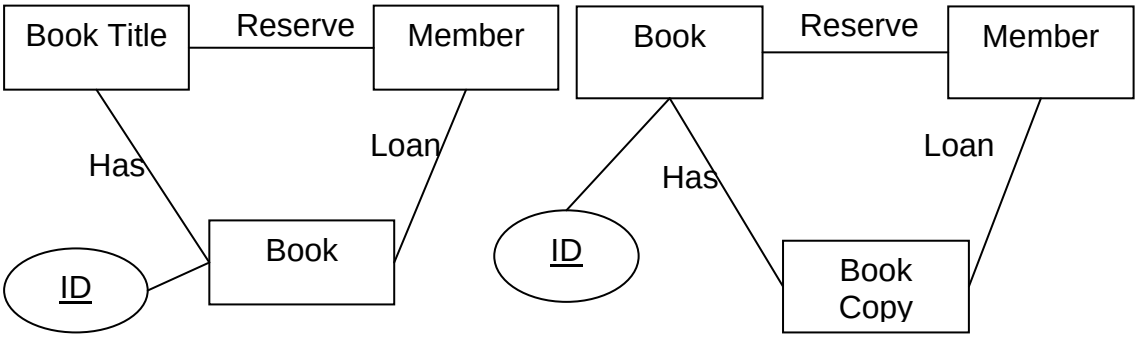
The aim of the semi-automatic assessment is to reduce the number of diagrams marked by the assessor. The system groups identical segments of the student's diagrams and then asks the assessor to approve the correctness of a diagram fragment from the each of the different groups. Therefore the assessor would be involved in the marking process only for the number of diagram groups rather than the total number of student diagrams.

Grouping the diagram pieces not only reduces the marking load but also makes the marking process consistent. The assessor doesn't have to repeat their judgement on the identical diagram pieces from student diagrammatic solutions. This repetition may lead to inconsistency in marking. The approved groups can be automatically graded easily and consistently by the system. Therefore grouping correctly is the key part of the system that enables the system to provide standardised marking.

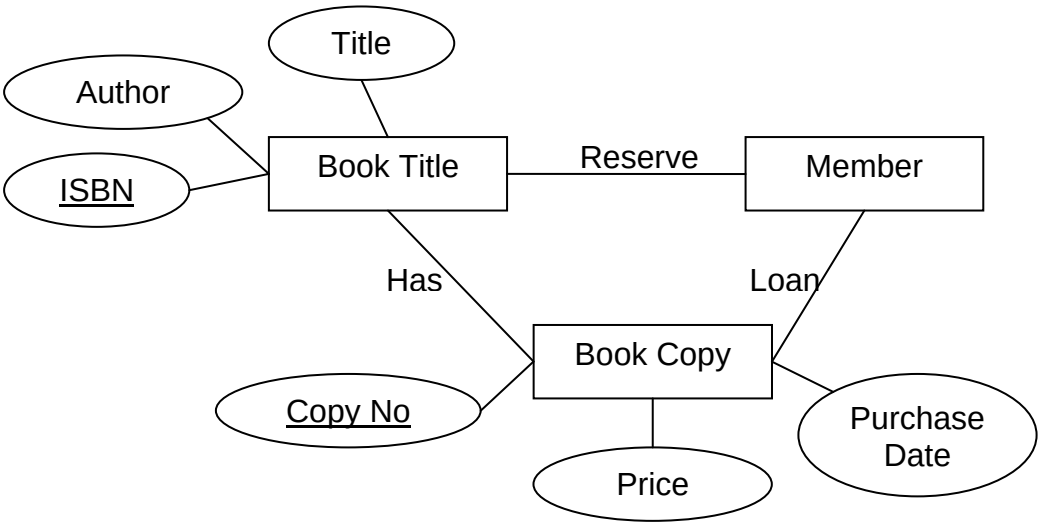
The correctness of the grouping depends on the criteria used to match the diagram pieces. The smallest diagram piece in each group can be either an entity or relationship for a conceptual database diagram. Entities in different diagrams could be considered as matched exactly if they have the same name and the same number of attributes with same name. This initial definition is pretty tight and finding two identical entities among student diagrams may not be trivial. This would increase the number of times the assessor is involved to decide whether the fragment is acceptable or not. However, it might be argued that if the same question is asked many times over the years then it can still be beneficial. Even if we accept this argument, grading a new student diagram by matching previously marked diagram fragments may not work correctly in some cases by using this matching criterion.

The diagrams in figure 1 belong to part of two different student diagrams based on a same scenario. "Book" entity in the first diagram clearly corresponds to "Book Title" with the missing attributes in the teacher solution. However, "Book" entity in the second diagram corresponds to the "Book

Copy" entity. The tool would not get the assessor to mark the second "Book" entity since it matched with the previously accepted "Book" entity by giving it the wrong meaning. Therefore, even the tight definition above is not sufficient for correct entity matching. The definition should also include contextual attributes of an entity. On the other hand, increasing the number of matching criteria required is counter productive.



Student 1 Diagram Student 2 Diagram

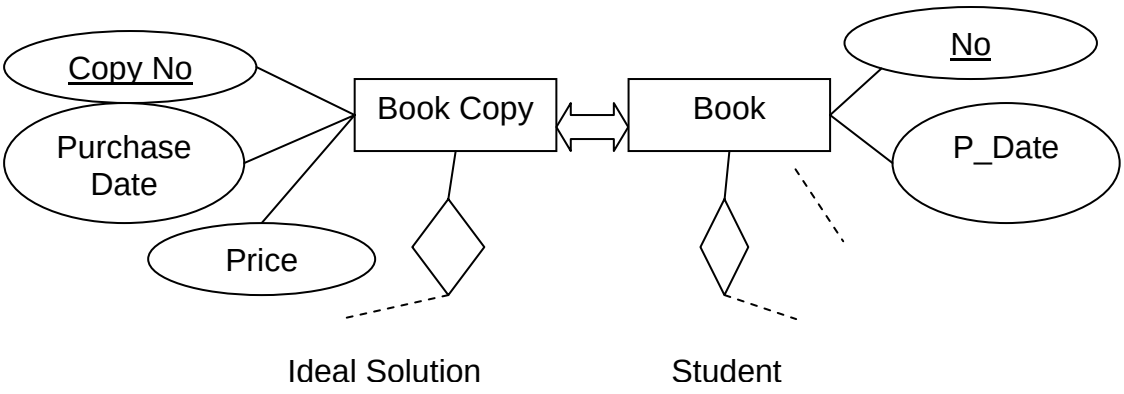


Teacher Diagram

Figure 1. Entity Name Ambiguity

The DEAP Project uses Latent Semantic Analysis (LSA) in order to determine the context of each diagram component. LSA semantically matches a word between the student text and teacher text by means of a factor analysis [4]. It relies on a large corpus of texts to build a high dimensional semantic space containing all words and texts. For instance, the word bike occurs generally in the context of handle bars, pedal, ride, etc [7]. Therefore, if a word like bicycle occurs in a similar context, the two words will be considered close the each other from a semantic point of view. The DEAP Project have recently reported that two small quite different diagrams can be regarded as equivalent [8], which is a result of using LSA. LSA doesn't work properly in the essay marking if the text size is small [7]. The DEAP Project are currently trying to overcome this problem.

In the KERMIT approach contextual meaning of an entity is given by explicitly forcing the students to highlight the related text in the scenarios. This approach simplifies finding a semantic match of the two components automatically (in figure 2). However, finding a related text to diagram components is not a straightforward task [9] and also the direct correspondence sometimes doesn't exist. The main reason is that designing a conceptual database model is an iterative process. Although the initial diagram can have a direct link to the scenario text, afterwards that initial diagram is subject to modification by applying designs rules and constraints in the domain. Although the final diagram can have implicit links to the scenario text, it is not always possible to show those links explicitly without all the intermediate steps between the initial and the final diagram (figure 3).



--- --- ---- ----
Each **book copy** has got a unique copy number and their price and purchase date is recorded.

Figure 2. Entity matching in KERMIT: Book copy and book entity are same concept

Our research suggests using not only the reference text but also the intermediate diagrams in order to define the contextual meaning of a component. However, not all intermediate diagrams are important for the context. For example, a student could initially consider the "book copy" noun phrase in figure 2 as an attribute of an existing entity in their diagram and later on they could change the attribute to an entity. It is not important to know this step to identify that component of the diagram. However, in the case that the student merges "head of department" and "lecturer" entity type to create "staff" entity type (see figure 3), knowing these intermediate diagrams is necessary to be able to match diagram components. We will call the former a direct referenced (DR) component and latter an indirect referenced (IR) component. This research proposes a tool to record the previous diagrams leading to the IR-component only.

The intermediate diagrams used for the contextual meaning also represent students' reasoning process during the design. When the assessor is presented with the intermediate diagrams of each component group for

marking, they can see the process of the students' thinking that enables them to give accurate feedback to students. However an extra caution should be taken not to overwhelm the assessor with so much diagram information during marking. Later our research will investigate how best to display the diagrams to the assessor.

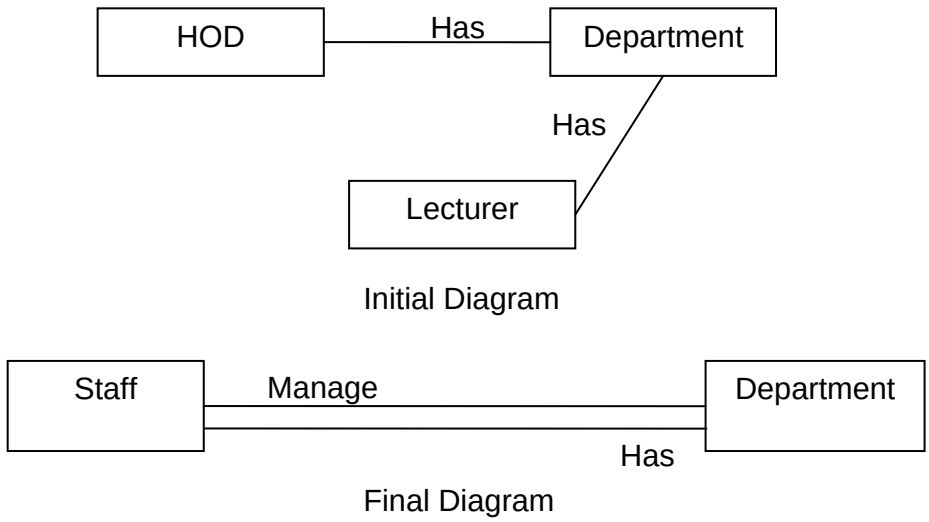


Figure 3. Conceptual Database Design is an iterative process

Merging two entities is one of the diagram modifications which results in IR-components. When the student decides to merge two existing entity types in the diagram, they could modify the diagram in various ways. For example, they might remove those entities and create a new one rearranging all the attributes and relationships of those entities or they might remove one of the entities and rename the other entity. After that, they identify the attributes and relationships of the new entity. These student actions must be interpreted to be able to identify a merging event. Even then, the interpretation may not be what the student intended. We suggest that the student needs to explicitly mention their intention during the design. This method is called self–explanation in the literature [10].

Psychological studies [12, 13] show that self-explanation (SE) is a very effective learning strategy resulting in deep knowledge. SE systems support students while they study solved examples or are asking for an explanation while solving problem. The main problem of self-explanation whilst solving the problem is the high cognitive load [9]. The proposed diagram editor is designed to reduce the cognitive load of self-explanation. The next section looks at components of this diagram editor and examines how cognitive load may be reduced.

Diagram Editor

The prototype diagram editor is based on automatic graph drawing [11]. The editor is an environment to capture student database designs. It is believed

that the student shouldn't have to draw a diagram for their design. They would simply enter the component type and name and then the tool would draw the student diagram. In this way, they can focus more on designing than drawing.

It is also believed that the automatic diagram drawing has advantages over the normal drawing tool in assessment. For example, analysis of database exam scripts reveals that students often change their diagram during the design. Moreover, some of them redraw the whole diagram when they have finalised the design. The automatic drawing could save student time during the exam in this case. Additionally, Thomas [8] found some evidence that the different orientation (shape) of identical student diagrams could be graded differently. The inconsistency of the marking can be prevented by the automatic drawing tool since it always draws the diagram in the same shape for an identical design.

The prototype editor consists of three sections; scenario text, diagram display and diagram modification sections. The scenario text section shows the scenario paragraph by paragraph so that the student considers the information in that section only. This method is called scaffolding in the self-explanation literature [12]. This section also has a feature to highlight the referenced noun phrase and sentences for the selected component. As for the diagram display section, it simply shows the automatically drawn ER-diagram of the student design. In the prototype the database diagram is not drawn or refreshed until the "Draw" button is pressed.

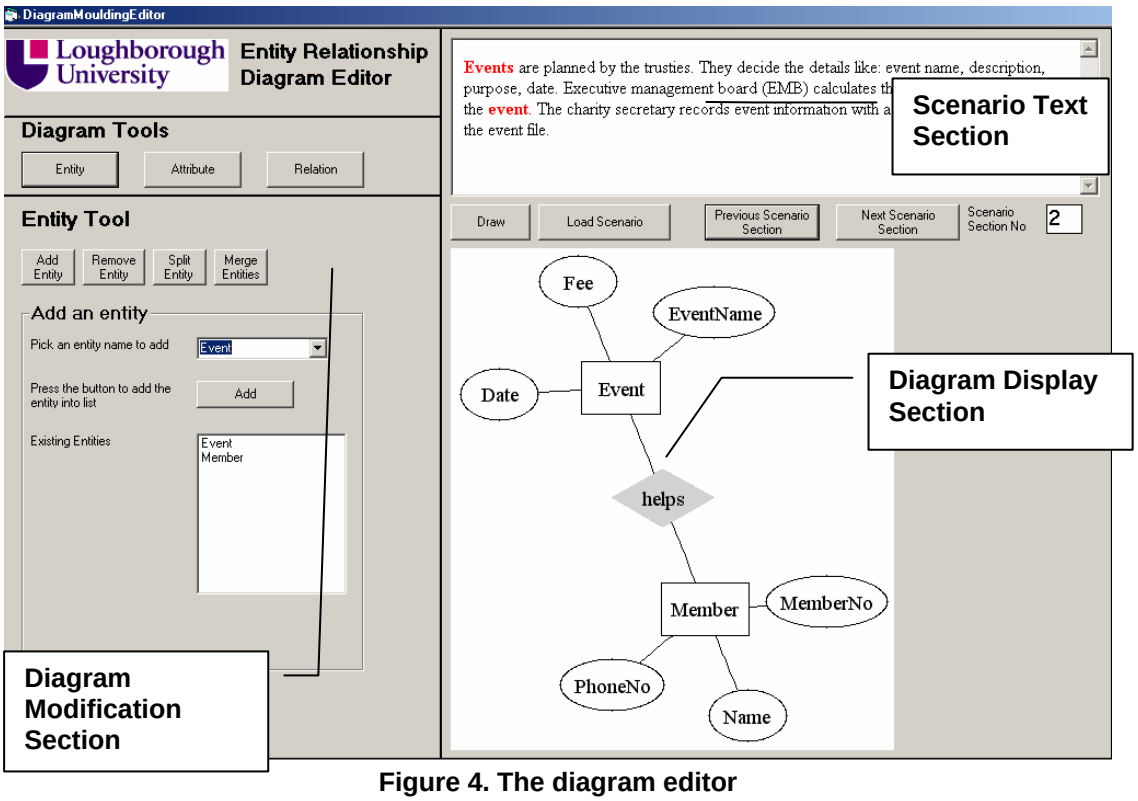


Figure 4. The diagram editor

The diagram modification section is the main part of the editor. In this section the student can add new components or modify existing ones. To create a

new entity type or attribute the student picks the component name from a list. The list has got all different noun phrases appearing in the current paragraph of scenario text. In this way, direct reference of the component is captured. Unlike KERMIT, the editor does not allow the student to name the DR-component. It is believed that the naming sometimes causes inconsistencies between student diagram and the referenced phrase. For example, the student can highlight “member” noun and name “book title” to create an entity type. KERMIT also forces the student to highlight the noun phrase in the text rather than picking it from the list. The “picking” method is suggested to reduce cognitive load without losing any educational proprieties of the assessment. However, research is needed to compare the “highlighting” with the “picking from list” methods.

The student can also modify the diagram by changing existing components. The editor provides function buttons to apply this modification on components. For example, to split an entity into two entities, the student presses “split function” button and then fills the required fields. These buttons reduce the cognitive load of self-explanation.

Database modelling is an iterative process [9]. Students produce their design incrementally for the system. Students start the design with an initial diagram by identifying entities from noun phrases and identifying relationships from verbal expressions. Then they apply the design rules and system constraints to build their design until it satisfies all the system’s requirements. This conceptual database design methodology is supported in the editor. “Scenario scaffolding”, noun phrase list for each section and “Function buttons” are the important features of the editor forcing students to design their database model systematically

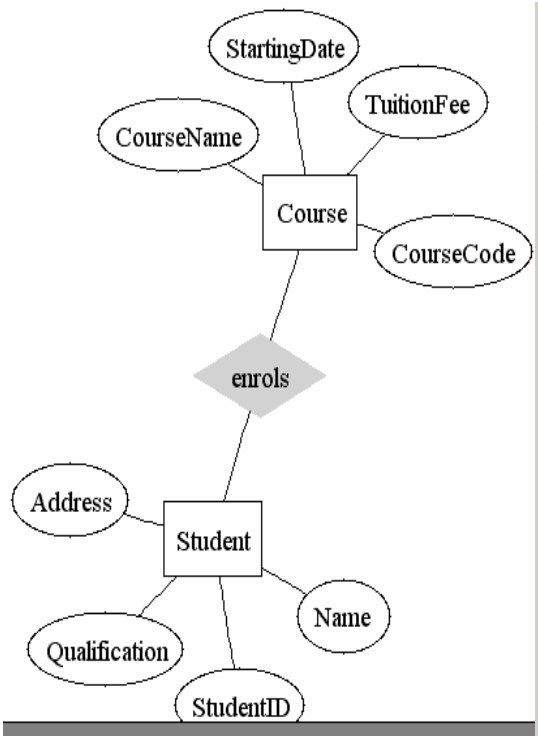
"The Blue Computer Training School *(BCTS)* provides a wide range of computer training short courses. *BCTS* 's manager gives you the following description of the business:

- The administrator records the details of any new course: course code, course name, description, level, tuition fee, and starting date.
- The details of new students are kept into the student file. The school needs to know their name, address and qualification. Each student is assigned a unique student id.
- A student may enrol on several courses. At the end of a course, the student is assessed and the grade achieved is recorded.
- Same course is offered several times a year. Students select a suitable starting date of the course during the enrolment.
- If necessary, the tuition fee of the same course is adjusted whenever it is offered. "

Figure 5. Sample Scenario Text

The scenario test in figure 5 requires using the “split” function button during design. The editor displays each bullet point of the scenario separately. The user sees the list of noun phrases which are in the current bullet point. Then they select a noun phrase to create an entity or an attribute of an entity.

Figure 6 shows an intermediate diagram of a user for this scenario. When the user considers the last two sections of the scenario, they may modify the diagram. The user needs to apply the “Split” function button for this modification (Figure 7), they then create relationships between “Course” and “Course offering” entities to reach the final ER-Diagram (figure 8).



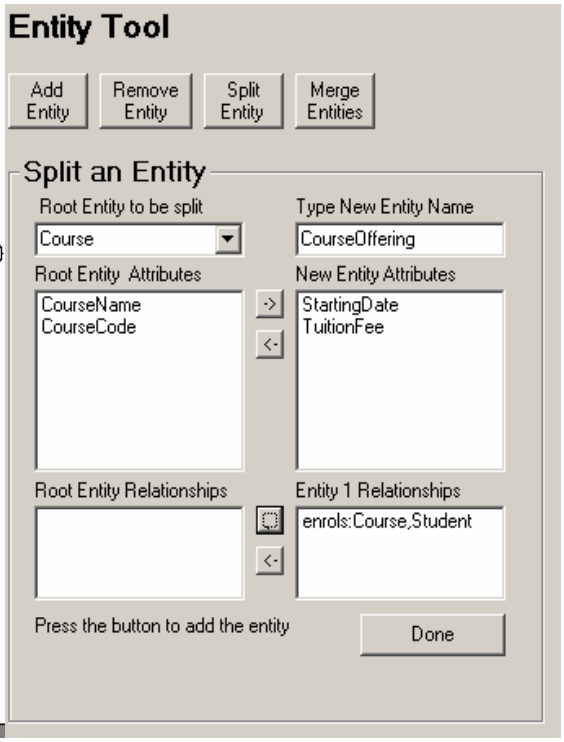


Figure 6. The intermediate diagram **Figure 7. "Split" function button box**

The tool is designed to have function buttons for diagram modifications which result in IR-components. However, function buttons for other kinds of modification can be also created. For example, changing an attribute to an entity type can be done by using function buttons. In this way, eventually, the reasoning processes of students can be gained as well as their final diagram. The examiner is able to understand student behaviour better and give more

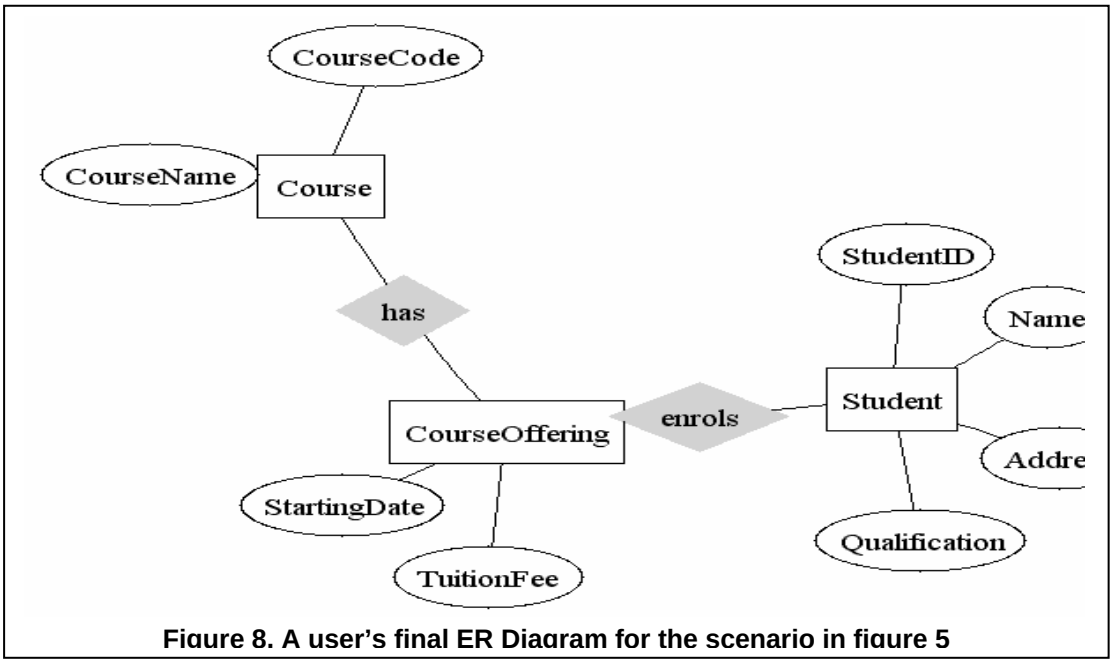


Figure 8. A user’s final ER Diagram for the scenario in figure 5

detailed feedback. On the other hand, there should not be too many function buttons since it increases the cognitive stress.

The usability of the diagram editor depends on the way the scenario text is written. If the scenario text is written in such a way that all the diagram components of the teacher's ideal solution are explicitly mentioned, then function buttons will not be needed. On the other hand, scenario text can be written in such a way that the student has to use function buttons to express their design or using the function button makes the design easier.

Experiment and Results

The diagram editor has two aspects. The first aspect of it is to capture contextual meaning of diagram components. This would help the examiner during marking. The second aspect is to provide an environment for the student to enter their design. Because of these aspects, the editor has a very different environment from those of traditional diagram drawing tools.

The users chosen for the experiment were people who have studied database design at university level. They were given an introduction session and shown how to use the editor on one example database scenario. The given example scenario uses one of the function buttons. Then the users are asked to design a conceptual database diagram for a similar scenario.

All the participants managed to draw the correct diagram. Although the given scenario didn't allow them to design the diagram without using the function button, none of them failed to use the editor. They all applied the required function button to modify the initial diagram during the design.

The required function button for the design expects an entity name from the user. All participants named the entity differently as expected. Different names for the same entity are not a problem for our approach since contextual information of the component is the main criteria for the entity match and this context is provided by use of the function button.

Conclusion and Further Work

The research investigated semi-automatic assessment which helps the assessor by reducing the number of diagrams to be marked. This paper proposes a new diagram editor which alters the traditional diagram drawing in order to make the assessment process suitable for semi-automation. This alteration removes the ambiguity of the contextual meaning for each component during marking. It also enables the assessor to better understand the student thinking and give accurate feedback to students. The prototype editor provides an environment in which the students can design the database model methodically and self-explain their design.

The editor was tested and initial results are very encouraging. They show that by using this editor the student design and contextual meaning of each design component can be captured without increasing the cognitive load on the student. However, further experiments are needed. Types of user and scenario are main factors which could affect the results. The users chosen could be students who are learning about conceptual database design, rather than experienced designers, and the given scenario could be written in such a way that it enforces the use of different combinations of the function buttons. Further experiments will only be done after completing the prototype editor. Currently the tool only has basic function buttons for a particular scenario type. All function buttons for different scenario types will be implemented.

The prototype has not focused on the “ease of use” aspect of the editor so the Interface needs to be made more user-friendly before the editor is used by students.

The other part of our semi–automatic assessment is the marker environment. The editor is a beneficial tool only if the contextual information of each component can be used by the marker environment to match them correctly. Therefore, implementation of this environment and experiments on it are also very important to complete the research.

Initial results with experienced database designers suggest that the tool is useful for designing database diagrams in their professional lives. This is not a current focus of the work but may become more important later on.

References

DEAP Project, Open University, LTSN-TLAD 2004 talk

Tselonis C Sargeant J McGee Wood M (2005). Diagram Matching for Human-Computer Collaborative Assessment Proceedings of the 9th CAA Conference, Loughborough University.

Suraweera, P., Mitrovic, A: KERMIT: a Constraint-based tutor for database modelling. Proc ITS'2002, LCNS 2363, 2002,377-387.

Landauer, T. K., Foltz, P.W. & Laham, D. (1998) an introduction to Latent Semantic Analysis.

Tsintsifas A., A framework for the computer-based assessment of diagram-based coursework, PhD thesis University of Nottingham UK,2002

Brett Bligh, Automatic Assessment of Diagrams, Feasibility Report, University of Nottingham UK, 2002

P. Dessus, B. Lemaire and A. Vernier, "Free Text Assessment in a Virtual Campus", Proceedings of the 3rd International Conference on Human System Learning, Europia, Paris, 2000, pp. 61-75.

P. Thomas, Comparing machine graded diagrams with human markers: some observation, Technical Report No 2004/27, DEAP Project.

Suraweera, P. An Intelligent Teaching System for database Modelling, MSc Thesis, 2001.

Chi,M.T.H. , M. Bassok, M.Lewis, P. Reinmann and R. Glaser, Self-Explanations: How students study and use examples in learning to solve problems. Cognitive Science, 1989. 15:p.145-182.

Graphviz is open source graph visualization software: http://www.graphviz.org

A, Bunt, C Conati, and K, Muldner, "Scaffolding Self-Explanation to Improve Learning In Exploratory Learning Environments", ITS 2004, LNCS 3220.

AN EVALUATION OF THE FORMATIVE FUNCTIONS OF A LARGE-SCALE ON-SCREEN ASSESSMENT

Andrew Boyle

An Evaluation of the Formative Functions of a Large-scale On-screen Assessment

Andrew Boyle
Research and Statistics team
Qualifications and Curriculum Authority (QCA)
83 Piccadilly
London
W1J 8QA
0207 509 5349
BoyleA@qca.org.uk
www.qca.org.uk

The key stage 3 (KS3) information and communication technology (ICT) test is an on-screen assessment that is being developed by the Qualifications and Curriculum Authority (QCA) under contract to the Department for Education and Skills (DfES). It is intended that this test will be run on a statutory basis from 2008; providing a summary of every child's attainment in ICT at the end of the lower secondary phase on schooling.

The central output from this test is a national curriculum level for each pupil. However, the test also has a formative function; familiarisation and practice materials are available for teachers and pupils, and a formative report is generated for each pupil who completes two 50-minute practice test sessions.

This paper will report evaluations of formative aspects of the KS3 ICT test: findings from the literature into the formative use of e-assessment will be briefly reviewed. This review will contrast the key themes of researchers into e-formative assessment and those who are concerned with 'plain' formative assessment. This will, in turn, illustrate differences in approaches to formative assessment between secondary and tertiary education.

Next, the paper will report on opinions about the formative reports. Reported opinions will come from several sources:

- Minutes of Teacher Review Group and National Stakeholder meetings
- Findings from wide-scale questionnaire surveys
- Findings from a survey conducted by telephone interviews

The paper will conclude by stating a summary evaluation of the KS3 ICT test formative reports. It will also go further to consider the implications of the findings with respect to the KS3 ICT test for the wider use of e-assessment for formative purposes.

THE SPRINTA PROJECT: ENGAGING STUDENTS IN FORMATIVE ASSESSMENT: STRATEGIES AND OUTCOMES

Nicola Bryan, Guinevere Glasfurd-Brown and Martin Sellens

The SPRInTA Project: Engaging Students in Formative Assessment: Strategies and Outcomes

Nicola Bryan, Guinevere Glasfurd-Brown**,** Martin Sellens
University of Essex
01206 874369
njbrya@essex.ac.uk
guin@essex.ac.uk
sellm@essex.ac.uk

The SPRInTA Project, (Student Portal Resources for Innovative Targeted Assessments), is a two-year project at the University of Essex funded through the Higher Education Funding Council for England (HEFCE) as part of Phase 5 of the Fund for the Development of Teaching and Learning (FDTL). The aim of the project is to provide a solution to sector wide concerns that increasing student numbers and unfavourable staff to student ratio's are adversely affecting the support available to students on assessment and the provision of effective feedback. The project aims to address this issue by developing tutorial guidance and formative assessments for undergraduate Sports Science Students. These resources are being made available via the University's institutional student portal, enabling targeted support for assessment.

This short paper will provide a brief overview of the SPRInTA Project detailing progress to date. Particular detail will be given to patterns of student uptake including strategies and recommendations for optimising student engagement in online formative assessment**.**

Introduction: An Overview of the SPRInTA Project

The SPRInTA Project is located at the University of Essex in the Centre for Sports and Exercise Science. The SPRInTA project is a two year project that started in November 2004 that aims to support student achievement by providing targeted and personalised support for assessment. The project is based on research that Computer Aided Assessment has numerous advantages, especially when used for large groups of students and can be used to give students' feedback, guide student effort, diagnose problems in learning and can give students experience in assessment methods (Lowry, 2005).

Over the two-year period the project team have developed a range of online formative assessments and tutorial guidance for Sports Science students. These formative assessments and tutorial guidance are based on the five

types of assessment common to Sports Science students, with the aim of achieving a high degree of transferability from the onset. The types of assessment used by the SPRInTA Project when delivering the project include multiple choice questions (MCQs), practical coursework, examinations, reporting in scientific paper format (SPF) and data analysis and interpretation.

Unique to the project is the automated delivery of dynamic assessments which are made available to students via the myEssex student portal[i]. When a student logs into the portal they will be recognised and will be served assessments and learning resources that relate to their record of achievement at that time. In this way the project enables tailored learning pathways and will in effect deliver 'intelligent' assessments. As the University of Essex uses QuestionMark Perception extensively to deliver computer assisted assessments this will be the software of choice for the SPRInTA Project and will also use the related programming tool QMWise to develop this active link between the student and the online formative assessments.

Project Progress to Date

To date the SPRInTA Project has developed large stratified MCQ banks for a number of level one and level two modules. These modules include; Human Physiology, Biomechanics, Sport Psychology, Nutrition and Metabolism, Functional Anatomy and Exercise Lifestyle and Health. The question banks are designed as formative assessments and a way that students' can self-assess themselves online during the course of the module.

In the level one modules that SPRInTA have targeted, summative assessment is via an MCQ examination at the end of the module and by a short answer and essay examination in the summer exam period. By providing large MCQ banks for student self-assessment a high degree of transferability from formative assessment to summative assessment was available from the onset. This was further developed by a switch from paper based summative MCQ exams to online summative MCQ exams. The consistency in the format of assessments has deemed to be very popular with level one students.

Questions were authored in QuestionMark Perception for each module and were then split into either submodules (for the pilot module) or weekly releases (subsequent modules after the pilot). The questions were then divided into three difficulty levels (basic, intermediate and advanced). On the last teaching day of each week (or subtopic) a set of questions relating to that weeks topic or submodule became available via the myEssex student portal. The student could then access the questions at an intermediate level. Once the student had submitted their intermediate test, they receive full feedback for each answer and feedback for the assessment as a whole, depending on their score they then got the chance to re-take the intermediate assessment (40-80%), or a more basic (<40%) or advanced (>80%) assessment.

The online formative assessments have been very popular with the students, a recent survey demonstrated that 94% of students recommend online self-assessment should be made available for all first year modules and 76% of students agreed that SPRInTA self-assessments have aided their learning. This improvement in learning has also been demonstrated by improvements in academic performance, statistical analysis (independent t-test) has shown that the introduction of online formative assessment resulted in significant ($P<0.01$) increases in the summative Multiple Choice Question (MCQ) exam results when compared with the results from previous years. This improvement has also continued into the Human Physiology summer exam with a significant improvement ($p<0.05$) in performance from the previous year. At the time of writing, it is too early to report on results of the additional modules that SPRInTA has developed, but it is expected that the improvement in summative MCQ exam scores will also be replicated in the summer exams.

SPRInTA are currently working on the second phase of the project, the tutorial guidance section. Interactive virtual learning environments using WebCT are being produced for reporting in scientific paper format (SPF), essay writing and data analysis and interpretation.

Strategies for Engaging Students in Online Formative Assessment

Gibbs and Simpson (2003) argue that 'you have to assess everything in order to capture students' time and energy'. The SPRInTA team were very aware that student uptake of formative assessment can be poor when formative assessment is "un-assessed". To try and avoid low student uptake the SPRInTA Team decided to provide an incentive to encourage students to use the formative assessments provided and the 30 end of module summative MCQ questions were placed within the question bank for each module. A minimum of three hundred questions were used for each module to prevent rote learning.

SPRInTA split the assessments into weekly or submodule releases, to encourage students to engage with the assessment on an even basis throughout the module. However, as previously reported (short paper presented to the 2005 CAA Conference; The SPRInTA Project: Supporting Student Assessment through a Portal) in the original pilot run by the SPRInTA Project on a level one Human Physiology module student uptake was heavily skewed towards the exam period with a significant majority of assessments being completed in the week before the final exam (see Fig 1).

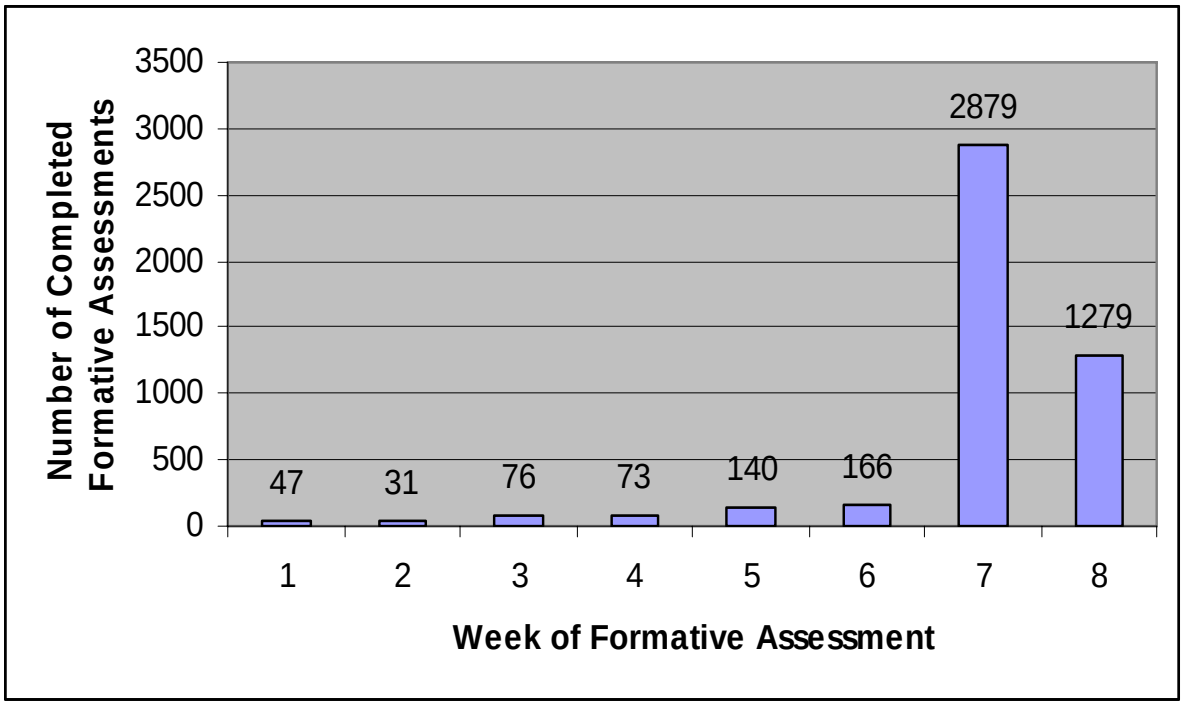


Fig 1: Number of formative assessments completed per week for the Human Physiology pilot

In order to try and modify this uneven completion of assessments a number of initiatives were implemented to subsequent modules to try and encourage students to spread their study time evenly over the course of the module. Many of these changes came from information collected in the survey and focus group that were completed at the end of the pilot module and the changes were implemented in October 05.

Feedback: Preventing Surface learning

It was decided to alter the feedback that was given in the pilot whereby the feedback to an incorrect distracter was the correct answer, as it was felt that just giving the correct answer encouraged surface learning. In order to engage students and encourage a deeper level of learning the reason why the distracter that was selected was wrong was given instead. Also in an attempt to give “correct guessers” more learning opportunities, feedback was also given for correct answers.

Feedback was also included within the myEssex student portal to enable users to view the date and time the assessment is available from and until, the number of previous attempts for each assessment, the maximum, minimum and average score for each assessment as well as their score for their previous attempt.

Publicity: Raising the Profile of the SPRInTA Project

In the survey at the end of the pilot 83% of students had heard of the SPRInTA Project and 73% of students thought the question banks were well publicised by SPRInTA. To try and improve the profile of the SPRInTA Project

every level one student was given a SPRInTA key ring at the start of term. The key ring was also a bottle opener and the theory was that each time the student used the key ring they would be reminded of the formative assessments available to them.

It was also decided to try and improve the way that students were alerted when new assessments became available. An additional slide was attached to each lecturers PowerPoint informing the students when a new assessment was available. The slide was designed to catch the student's attention and included a picture of the SPRInTA Project Officer with some speech attached regarding the new assessment.

Release Dates and Patterns

Data from the survey indicated students did not like the randomisation of questions because they could not guarantee that they had viewed and completed all questions when revising for their end of module summative assessment. The SPRInTA team decided to reduce the question bank to 300 questions (600 questions were written for the pilot module) and to release assessments on a weekly rather than sub-topic basis. This meant that each weekly assessment would contain 38 questions (12-13 questions for each difficulty level), and students who completed all assessments would have completed every available question.

The release dates and patterns of assessment were also experimented with. Two core level one modules (A = Functional Anatomy and B = Sports Psychology) were supported by a weekly release of online formative assessments. Module B was exposed to the weekly release pattern as previously reported and outlined above. This release pattern involved weekly topics of formative assessment opening throughout the module and staying open until after the summative assessment. Module A received formative assessments that were open for two weeks and then closed until the week before the summative exam, when they were again made available for revision purposes. The hope was that this would encourage a more even distribution of student participation throughout module.

Student Engagement and Feedback

As seen in the Human Physiology Pilot engagement levels with the online formative assessments was high, 83.7% of Module A and 79% of Module B students completed at least one assessment.

Statistical analysis (independent t-test) demonstrated that module A and module B saw a significant ($P<0.01$) improvement in the end of module MCQ summative exams when compared to the previous year. Module A also saw a significant positive correlation ($P<0.05$) between the number of completed formative assessments and end of module MCQ summative exam.

Despite efforts to encourage students to use the assessments throughout the module, there was no change in the distribution of completed assessments as

seen in Fig 1. However it is encouraging to see that the percentage of completed assessments in the week leading up to the summative exam were less in Module A (68.2%) and Module B (56%) when compared to the previous Human Physiology Pilot (88.6%). This may have been due to the increased publicity drive with the end of module survey showing that 97% of students have heard of the SPRInTA Project when compared to the previous 83% seen in the survey at the end of the pilot. 97% of the students also thought the questions were well publicised by SPRInTA (previously 73%), with 76% of students finding the new assessment PowerPoint slide useful.

The experimentation of release patterns had little effect on the distribution of completed assessments; this can be seen in Fig 2 and Fig 3.

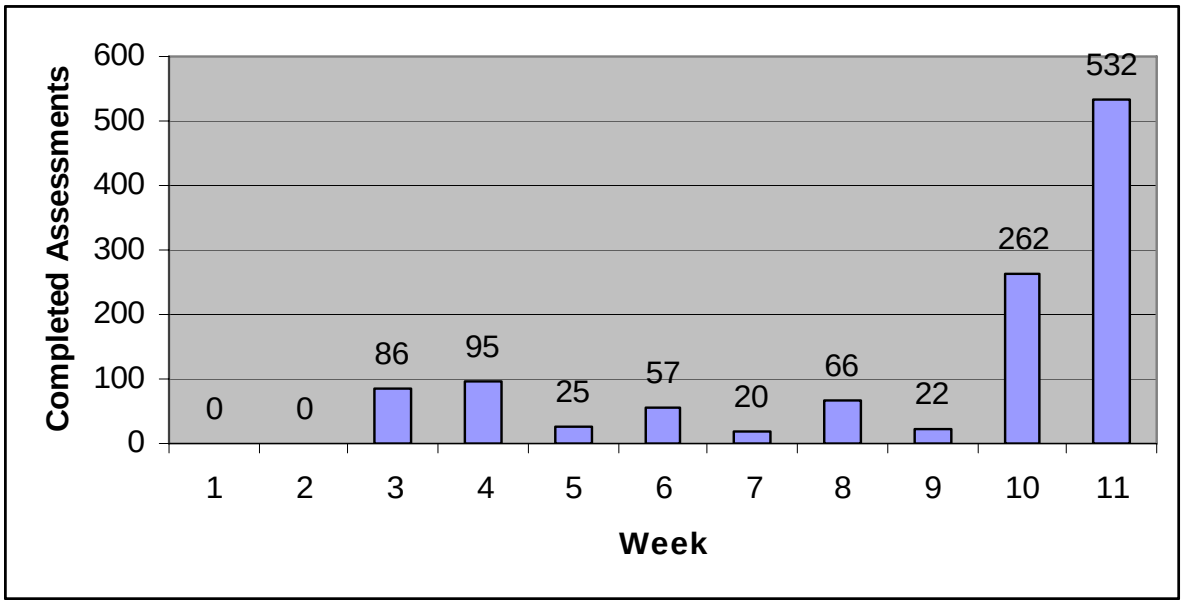


Fig 2: Number of completed formative assessments by week number for Module A. The Summative Exam was at 10am on the Tuesday of Week 11

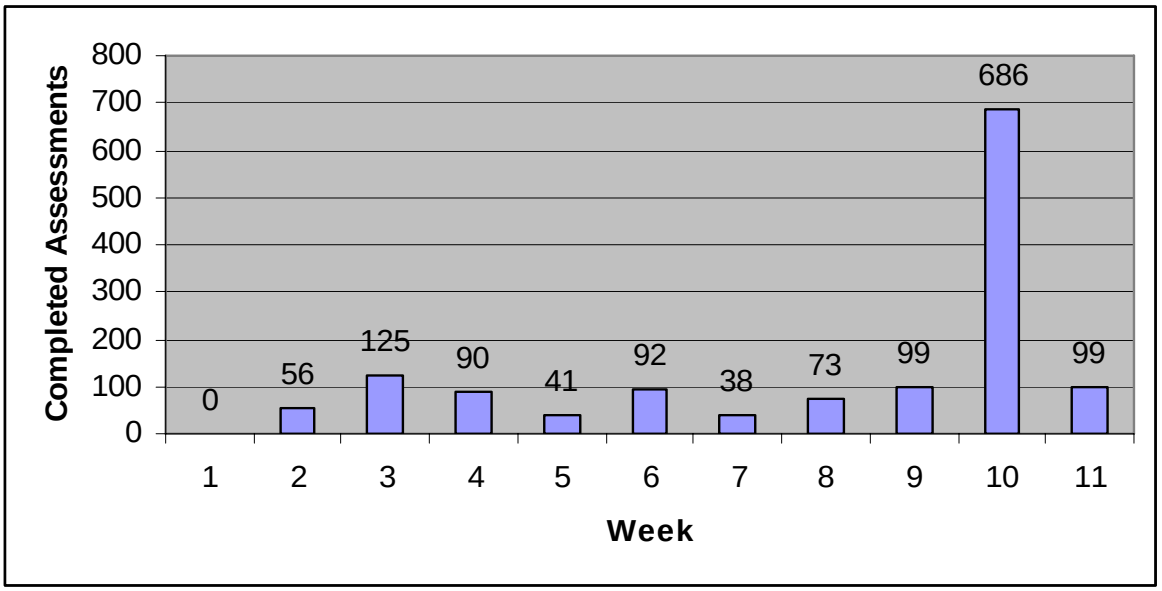


Fig 3: Number of completed formative assessments by week number for Module B. The Summative Exam was at 10am on the Monday of Week 11

In addition students disliked the release pattern implemented for Module A. In a survey completed at the end of module A and module B students were asked whether the new release pattern for Module A helped them to distribute their study time evenly, only 31% of students agreed that this was the case with 21% of students neither agreeing nor disagreeing and 48% of students disagreeing. During completion of the survey students were invited to give free text responses regarding the positive and negative aspects of the SPRInTA question bank. There were no positive comments for the release pattern trialled in Module A, in contrast 21% of the negative comments were about the release pattern.

Conclusion

It can be clearly seen that students like freedom of choice when choosing when to study, and despite attempts to alter study patterns it appears that a majority of students when revising for a summative MCQ exam are strategic/surface learners. This agrees with the some of the current literature available on MCQ tests which suggest that MCQs only measure the first level of intellectual behaviour important in learning (knowledge). A study by Scouller (1998) showed that students were more likely to employ surface learning approaches in the MCQ examination context and to perceive MCQ examinations as assessing knowledge-based (lower levels of) intellectual processing. In contrast, students were more likely to employ deep learning approaches when preparing their assignment essays. SPRInTA are in the process of collecting data from the students about their approach to learning to validate these claims.

It can be concluded that in order to engage students in formative assessment there needs to be an incentive for the student. In this case it was summative

questions placed in a large bank of formative questions; however there are some further ways that the SPRInTA Team are looking at engaging students.

Ranking

As the SPRInTA Project are working predominately with Sports Science Students it has been suggested that we appeal to their competitive nature and add a ranking system to the information that the student receives about their assessments in the student portal. This means that once a student completes an assessment they can see where they lie in terms of performance against their peers.

Summative Component

According to Tait *et al.* (1998) the strategic approach refers to the systematic arrangement of learning activities in order to achieve the specific assessment criteria required to pass a course. If the summative component of the course is at the end of the module (as seen in Module A and B) this means a strategic student will only study in the lead up to the exam (as seen in the SPRInTA initiative). If a small summative component was attached to each weekly assessment this would make the assessments compulsory and as a consequence build a more consistent and deeper approach to learning. In a recent focus group this was deemed popular with the majority of students as less pressure would be placed on the student in the end of module exam.

Just in time Teaching (JiTT)

Just-in-Time Teaching is a teaching and learning strategy based on the interaction between online assessments and an active learner classroom. Students are required to complete an online assessment before a lecture, before the lecture the lecturer reads the student submissions "just-in-time" to adjust the classroom lesson to suit the students' needs. This could work well with the SPRInTA Project as it would ensure consistent engagement in with the question banks as well as tailoring lectures to the students needs.

Further information about the project can be found at

http://www.essex.ac.uk/sprinta/

References

Gibbs, G. Simpson, C. (2003). Does your assessment support your students learning? *Learning and Teaching in Higher Education*, **1**.

Lowry, R. (2005). Computer aided self assessment-an effective tool. *Chemistry Education Research Practice.* **6** (4), 198-203.

Scouller, K. (2006). The influence of assessment method on students' learning approached: Multiple choice question examination versus assignment essay. *Higher Education.* **35** (4), 453-472.

Tait, H., Entwistle, N.J., & McCune, V. (1998) 'ASSIST: a reconceptualisation of the Approaches to Study Inventory', in C. Rust (ed.) *Improving Students as Learners.* Oxford: Oxford Centre for Staff and Learning Development, Oxford Brookes University

[i] The myEssex student portal offers students structured sets of links to online services and information, customised for each user, and with further options for users to personalise a range of features. The portal delivers customised links and information based on what it knows about the user (you are studying these courses, you are based at Loughton/Colchester, etc) and personalised by the user (the user can choose to hide some links, add others, and change the presentation).

ACCEPTANCE AND USAGE OF E-ASSESSMENT FOR UK AWARDING BODIES – A RESEARCH STUDY

Geoff Chapman

Acceptance and Usage of e-Assessment for UK Awarding Bodies – a Research Study

Geoff Chapman
Thomson Prometric
www.thomson.com
Geoff.Chapman@Thomson.com

Abstract

This research provides an exploration of the UK e-Assessment market, in relation to the UK Awarding Bodies, comparing findings with those of twelve months ago. It also elucidates on the key areas that have emerged since the first research was conducted.

This provides an insight into the remaining drivers and barriers to the adoption of e-Assessment, but also the widespread acceptance and adoption in the UK.

With 81% of all recognised Awarding Bodies being interviewed, this study is verging on an Awarding Body e-Assessment census based on sound research principles which will lead to continuing e-Assessment development.

The level of e-Assessment industry knowledge and uptake of programs within UK Awarding Bodies is at a much more advanced position compared to the previous research findings. The pace of market change has clearly quickened. It is possible to state that these findings will allow Awarding Bodies to revisit their thoughts on e-Assessment, altering the pace of market maturity in the short to medium term.

Questions related to topics such as psychometrics, use of multiple choice questions for higher levels of learning and e-Assessment location preference, have provided responses which give a sign-post for the key emergent market needs.

Overview

Using the previous study as a benchmark and noting the changes in the regulatory environment and further exploration of e-Assessment issues by the QCA, it was decided to consider the acceptance and usage of e-Assessment.

The new study would once again consult the QCA recognised UK Awarding Bodies and other key stakeholders as to their level of acceptance and usage of e-Assessment. Within this market, there are a handful of UK organisations,

outside of the UK Awarding Body field, having great e-Assessment experience and using ‘mature systems.’ Similar to the previous study, it was important to capture the input of these organisations at the qualitative stage, so that the quantitative phase could be as fully informed as possible.

The idea of a ‘census’ of the 115 UK Awarding Bodies recognised by QCA was retained with the contacts being those who have a specific responsibility for their organisation’s exam or qualification system.

Number of Quantitative Respondents

93 respondents from 115 Awarding Bodies (currently accredited by QCA) responded to the quantitative phase of the research. This covers 81% of the research universe. This exceeds the 87 respondents from 116 Awarding Bodies from the previous study.

Key Findings

The headline finding from this study is that 38% of Awarding Bodies surveyed currently use e-Assessment to deliver up to 60% of their assessment programme. If the rate of change remains the same, e-Assessment will soon be adopted by over 50% of Awarding Bodies: a clear majority. e-Assessment has now achieved ‘acceptability’ within the marketplace with strong majority verdicts on understanding, acceptance and usage.

The key benefits of current e-Assessment are now being understood as more organisations implement the changes. Market movement and increasing recognition of factors such as ease of administration and time flexibility are hallmarks of systems that have successfully been bedded into organisations and accepted by stakeholders.

An area of business concern that was raised in the qualitative phase was the notion of e-Assessment’s return on investment (ROI). Clearly when significant resources are staked in e-Assessment, stakeholder interest in delivering organisational benefits are paramount. Seven out of ten respondents believe that e-Assessment will deliver ROI – clearly a sign of confidence in how it can improve not just the candidate experience, but also deliver efficiency savings and / or stakeholder value.

The subject of psychometrics was also flagged in the qualitative phase as being one of emergent, but increasing importance. Whilst there is limited understanding of the subject at large, this can be compared to the weak knowledge regarding item types that was highlighted in the previous study. If the positive results regarding multiple choice items is indicative of how the market can quickly assimilate e-Assessment knowledge, it would be reasonable to suggest that the market knowledge of the benefits of psychometrics in e-Assessment will rise quickly.

The focus on candidate needs is called out by a number of the findings. It is pleasing that whilst there is an acknowledgment of commercial factors,

candidate needs such as accessibility and time flexibility remain at the forefront. The importance of candidate satisfaction remains a key importance factor for Awarding Bodies.

The need to make the most appropriate and best e-Assessment choice is a suggestion arising from the strong call-out for multiple technologies conforming to agreed standards. As wider issues such as the Unique Learner Number and ID cards impact on facets such as registration, exam booking and candidate verification, the need for differing e-Assessment systems to have a mutually compatible interface point, recognised and mature e-Assessment standards will become more important. Additionally, this raises a flag to e-Assessment providers to ensure that their systems are capable of adhering to the demands of these standards.

Conclusions

The acceptance and usage of e-Assessment has clearly grown at a substantial rate compared to the previous research study. The strong confidence shown in the ability of e-Assessment to deliver return on investment is a major finding of the research. The use of psychometrics is emerging with some usage reported by Awarding Bodies. Multiple choice question usage for higher levels of learning and high stakes exams is more widely understood and acknowledged.

Disadvantages traditionally associated with e-Assessment such as cost and technical issues have decreased in importance as uptake has increased and technology has matured. In parallel, areas of risk previously thought to be inherent in e-Assessment (data security and technology in general) are not as prominent as areas which perhaps are not exclusive to e-Assessment. Candidate authenticity is a key issue called out in the findings. The needs and desires of the learner/candidate continue to be at the forefront for organisations wishing to adopt e-Assessment or already using an incumbent system.

GENERIC MODEL OF COMPUTATION FOR INTELLIGENT COMPUTER AIDED PROGRESS ASSESSMENT (ICAP)

Esyin Chew and Norah Jones

Generic Model of Computation for Intelligent Computer Aided Progress Assessment (iCAP)

Esyin Chew and Norah Jones
Centre for Excellence in Learning and Teaching (CELT)
University of Glamorgan
Wales
CF37 1DL
echew@glam.ac.uk
njones2@glam.ac.uk

Abstract

One of the major problems levelled at many traditional learning initiatives is that individual progress and performance are not well monitored and evaluated. This paper offers a model of computation for intelligent computer aided progress assessment and reports on a recent study which formulated a generic model (iCAP) from a prototype testing in a 4 months course. A walk through study for the course was carried out which was used to formulate an intelligent computer aided assessment system. As a result, a generalized model was designed which was used to determine the expected performance bank with various levels of difficulty (challenge levels), thereby ensuring that, if the test is randomized, levels of competence could be examined. Each individual result of the student (current performance level) is captured and stored in a progress file for self-reviewing by the student as well as by the lecturer for assessment and monitoring purposes. The benefits and limitations of iCAP are discussed at the end of the paper.

Introduction

One of the criticisms levelled at many traditional learning initiatives is that they are not effectively monitored and evaluated (Thorne, 2003). The importance of effective assessment feedback in student learning is recognised by The National Audit Office in their report “Improving Student Achievement in English Higher Education” (2002) which indicates that the poor quality of academic feedback is a key factor in contributing to student dropout. The Quality Assurance Agency (QAA) code of practice on assessment (2000) also makes clear the need for timely and consistent feedback. The indication that assessment feedback is a concern for students emerged in the results of the National Student Survey (2005) where universities in the UK were consistently rated by their students as being poor in feedback on assessment.

The conventional assessment methods of learners in higher educational institutions are for example, quiz, tests, examinations, assignments or

projects. The student's learning performance is assessed at a certain point in time usually towards the end of a course, as a result the individual's progress is difficult to monitor in the traditional classroom. The lecturer may be aware of each individual's learning progress in a smaller class size but this would be a great challenge when dealing with a large number of students.

In conventional assessment methods, the learner tends to obtain the current state of his or her individual performance in an authoritarian and reactive way, and without a traceable progress history. The pragmatic educationalist, John Dewey's influence has been a leading factor in the abandonment of authoritarian methods and in the growing emphasis upon learning through experimentation and progression (Jay, 2003). The learner's knowledge will grow alongside the self-initiative experimentation in the learning process. It is essential that this progress is fully captured and recorded throughout the course and any learner's performance measures are based on these. Inge (1919, p15) also defined,

… *the aim of education is the knowledge not of facts but of values. Values are facts apprehended in their relation to each other, and to ourselves. The wise man is he who knows the relative values of things.*

There has been a growing literature on the impact of computers on education but more recently there has been an interest in blended learning. The blended learning environment is designed to aid the learners with state-of-the-art information technology in addition to the traditional face-to-face classroom. It combines face-to-face instruction with computer-mediated instruction (Bonk and Graham, 2006). Blended learning represents a more diverse combining of a variety of approaches such as coaching by a supervisor, participation in an online class and case studies (Jones, 2006). This paper contributes to the literature on blended learning and focuses on feedback and assessment.

This paper presents an intelligent computer aided progress assessment model, namely iCAP, to satisfy the agenda mentioned above. The proposed assessment model which captured and recorded the learner's progress played a part in providing essential values that the learner not only relied on the final marks given in each test or examination but also the satisfaction from the advancement to an improved or more developed state. Moreover, the lecturer could easily trace the learner's progress history to evaluate the achievement of its overall educational aims. The iCAP model is generalised from a system prototype which was tested in a teaching module in a local university. It is able to identify each learner's performance and the progress of improvement or decline.

There are few current systems in the market which have been analysed and evaluated and not many e-learning system have been designed with an inbuilt progress report facility. iCap has been designed to fill this gap. The summary is described in the below table:

Features/Tools	English-at-home.com	E-Classroom	(ITF) Modules	PrimeLearning.com
Progress Report	No	No	Yes	Yes
Teaching Material	Yes	Yes	Yes	Yes
Quiz/Test Module	Yes	Yes	Yes	Yes
System Type	Web-based	Web-based	Web-based	Web-Based
Developed By	English At Home	Mind Leaders	Arizona State University, US	AMA
URL	http://www.english-at-home.com/	http://eclassroom.internettoolkit.com/cgi/signon.exe?text1=demo&text2=demo	http://elearning.asu.edu/ITF_Module/	http://elearn.primelearning.com/prime/PrimeLearnerHomePage.jsp

Table 1.1: Summary Table for Current System Comparison

2.0 Analysis and Design for iCAP

Assessment plays an important part in providing essential information on whether the student is on track as required by the lecturer. If one of the purposes of education is to help close the gap between actual and desired performance we must be able to define what that original level of performance was (Thorne, K., 2003). The expected **performance level (plevel*)*** may vary from one lecturer to another. However it must be first defined before the assessment process is carried out. In this study, the plevel scale defined by the lecturer who was conducting the course is from level 1 to level 6 and the desired performance level is in level 3.

The scale for **level of challenge (lc)** varies from one lecturer to another. The lecturer defined lc in this research as “easy”, “moderate” and “challenging”. It is important to identify the difficulties of individual question - lc in the questions bank. There are two methods to obtain the lc:

Determined by the lecturer: this method is timelier but subjective because questions which are determined ‘easy’ by the lecturer may be difficult from the learner’s perspective.

Determined by past students: this method is objective but it may be tedious and time consuming to gather the necessary data.

The research is based on method (2) discussed above. 150 questions from lesson one to nine in the course were identified based on the course material. The tests were distributed to 28 students who took the module. It was conducted to determine the *lc* for all the questions to be placed in the questions banks. Respondents are required to categories the *lc* for each question as “easy”, “moderate” or “challenging”. The analyses of the

respondents' comments are concluded in Table 2.1. The *lc* of a question is determined by highest votes from the respondents.

Scale Level of Challenges (*lc*)	Lessons in the Subject		
	Lesson 1-3	**Lesson 4-6**	**Lesson 7-9**
Easy	15	16	14
Moderate	22	20	21
Challenge	13	14	15
Total	50	50	50

Table 2.1 Example of the Summary for Questions' Challenges Level (*cl*)

All 150 questions were grouped and stored in the database according to the different levels of challenge showed in the table. The number of questions in a test was set by the educator and in this case **15** questions per test are defined. The *plevel* associated with the *lc* is designed in the below table:

Performance Level	Easy	Moderate	Challenge	Total
1	10	5	0	15
2	8	5	2	15
3	6	5	4	15
4	4	5	6	15
5	2	5	8	15
6	0	5	10	15

Table 2.2 Example of the Association for *plevel* and *lc*

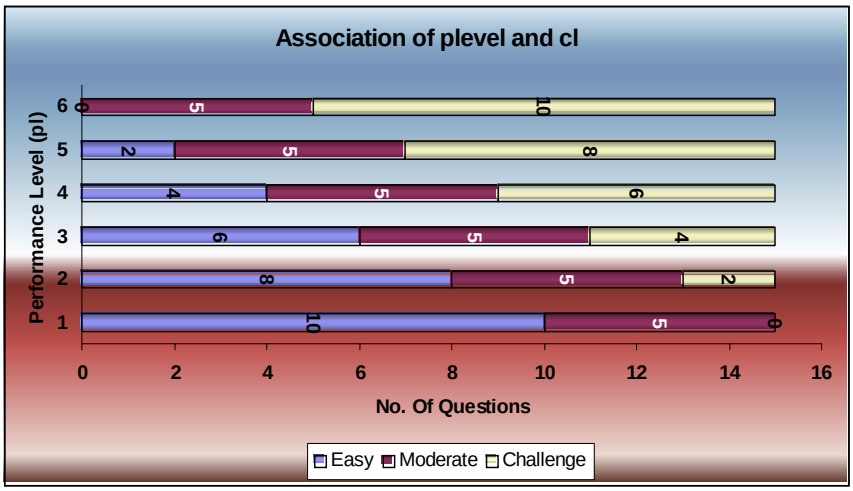


Figure 2.1 Association of *plevel* and *lc*

Figure 2.1 describes a phenomena that the higher the level of challenge *(lc)* is, then fewer the easier questions will be selected. Likewise, the higher the *lc* is, the more challenging questions will be selected.

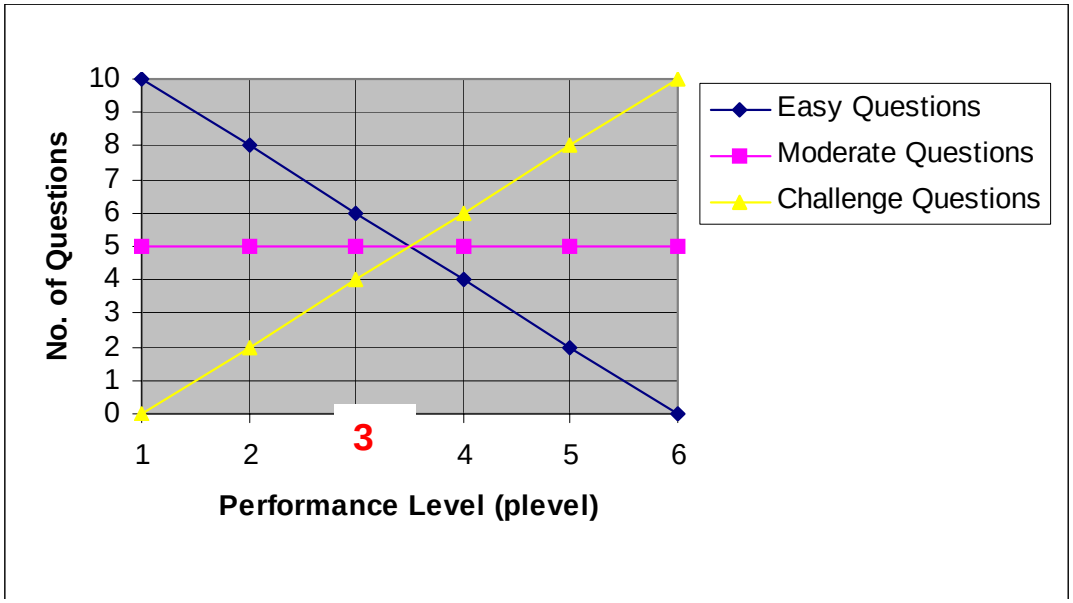


Figure 2.2 Desired Performance Level

Level 3 in Figure 2.2 is the predefined desired performance level. Each learner's default *plevel* was assigned to Level 3 upon their registration for the progress test, which contained 6 easy questions, 5 moderate questions and 4 challenge questions. There is a smaller set of performance level (*pl*) which determined the individual test if required. 60% score was defined as a *pl*. When the learner has completed the test, the level of challenge will automatically be decreased if the learner's result is below the desired performance level. Likewise, the level of challenge will be increased if the learner's result is above the desired performance level. Table 2.3 shows the scale of increase/decrease for *lc* in the study.

Score Range to *pl* in Each Test	**Level of Challenge to be Increased / Decreased**
90 - 100	+4
80 - 89	+3
70 - 79	+2
60 - 69	+1
50 - 59	0
40 - 49	-1
30 - 39	-2
20 - 29	-3
0 – 19	-4

Table 2.3 Example of Scale for the Increase/Decrease of lc

There will be no changes to the *lc* if the learner's current score is the same as the desired performance level (*pl*). For instance if the desired pl is now 50% and a learner's current score of the test is 58%, there will be no increase or decrease to the learner's current *lc*. Further elaboration is explained in the scenarios detailed below:

A learner is first enrolled in the course and the *plevel* is defaulted at level 3. The learner only manages to get 45% in the first test, which means that the learner fails the particular test. From table 2.2, the level of challenge is decreased to -1. The level of challenge in the next test will be set to (3-1) = 2.

If the same learner passed the second test with 85% score, the level of challenge is increased +3. The level of challenge for next test will be set to (2+3) = 5.

The maximum *lc* is at level 6 and the minimum *lc* is at level 1. Learners can only be assigned to the maximum or the minimum level even if their result required a level that exceeds the maximum or minimum level. Once the learner's level of challenge reaches the maximum level of 6 there would be no more increases. The same applies to the minimum level. Once the learner's level of challenge reaches the minimum level of 1 there would be no more decreases.

In summary, a set of predefined levels of challenge of test questions are generated randomly from the questions bank aligned with the learners' level of challenge history record. This means that the higher the marks scored by the learner the higher the level of challenge of the questions. In addition the learner's progress is captured and recorded for self-motivation and for the lecturer to monitor. Thus, this facilitates a unique subset of questions to be delivered for each assessment or each learner.

3.0 Generalisation of iCAP

The computation discussed above can be generalised into a generic model for intelligent Computer Aided Progress assessment (iCAP). First, the scale and desired performance level (*plevel*) are defined. The level of challenge (*lc*) is determined by probability samples (Cohen, Manion and Morrison, 2000) either by representatives of the sample (e.g. the lecturer) or a wider group of sample (e.g. students who has taken the course previously). The number of questions is identified and its relationship with *plevel* is showed in the below figures.

If y= Number of Questions, x= *plevel* and k = *lc*, the below graphs explain their basic association and relationship.

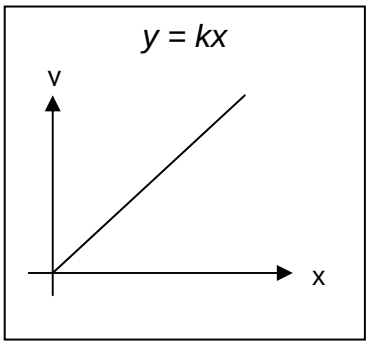


Figure 3.1 Challenge

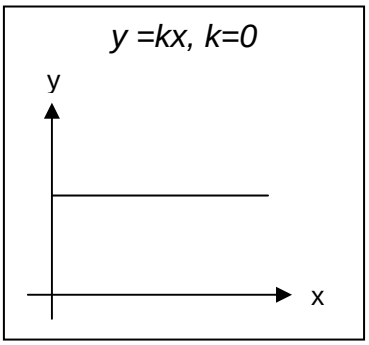


Figure 3.2 Moderate

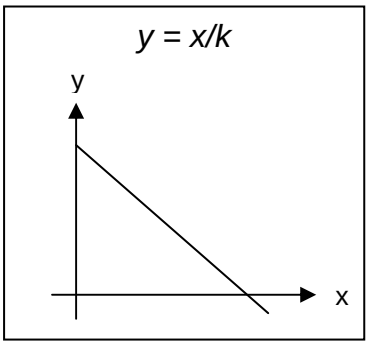


Figure 3.3 Easy

Figure 3.2 shows the default *plevel* or desired *plevel* which can be assigned to all learners when they first enrolled onto the course. Figure 3.1 illustrates the normal learning curve, which means the challenge questions which were selected from the questions bank are increased from level to level. Figure 3.2 shows the constant of the moderate questions when the level of challenge is increasing. Figure 3.3 shows how the easy questions decrease when the level of challenge increases. This simple model is used to formulate the test questions blended with its level of challenge as showed in the Table 2.2. Figure 3.4 depicts that when *lc* identified is increased or decreased (e.g.: *lc* = 5 {“Very Easy”, ”Easy”, ”Moderate”, ”Challenging”, ”Very Challenging”}). The educator can define the *lc* based on their requirements and preference for students’ assessment. The higher number of *lc*, the more complex the table of association for *plevel* and *lc* will be.

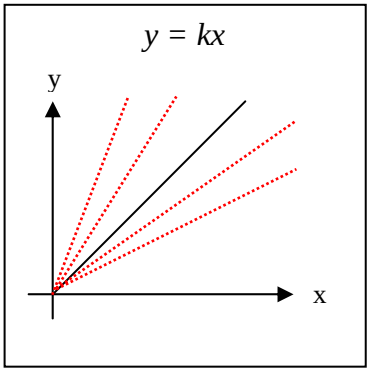


Figure 3.4 *y=kx*

The higher number of *lc*, the more complex the matrix table of association for *plevel* and *lc* will be. Thus, the generic table is:

plevel \ lc	*i*
j	Qs

Where,

plevel = Performance Level

lc = Level of Challenge

i = Scale value of *lc*, e.g.: easy, moderate and challenge.

j = Sequential value of *plevel*

Qs = Questions Selected

Qs is the number of questions to be selected in each matrix cell of *plevel* and each *lc*. It can be represented in the below computation:

$$(lc_1)_{plevel1} + (lc_2)_{plevel2} + (lc_3)_{plevel3} + \ldots (lc_j)_{plevel_i}$$

where i = 1 to total number of questions and j = scale of *lc*

3.1 Possibility for Repeating Question in iCAP

Each test consists of 15 objectives questions with 4 answer options. For each test, the database must consist of at least 50 questions. This is to ensure that the possibility for a single question to be repeated in the second set is lower or equal to 9%.

Possibility for a single question to be repeated in the 2nd set of question:

= 15/50 x 15/50

=1/2 x 1/2

=1/4

=0.09 or 9%

Figure 3.5: Possibility for a Question may be repeated in the 2nd set of Question

Although the possibility for a single question to be repeated in the 2nd set of question is 9%, which may be considered quite high, all the questions in the quiz are randomly arranged. This means that the possibility for a single question to be repeated as the same sequence in the 2nd set of questions is as low as $(1/50)^{15.}$

3.2 Generic Framework for iCAP

The research can be concluded in a generic model showed in Figure 3.6.

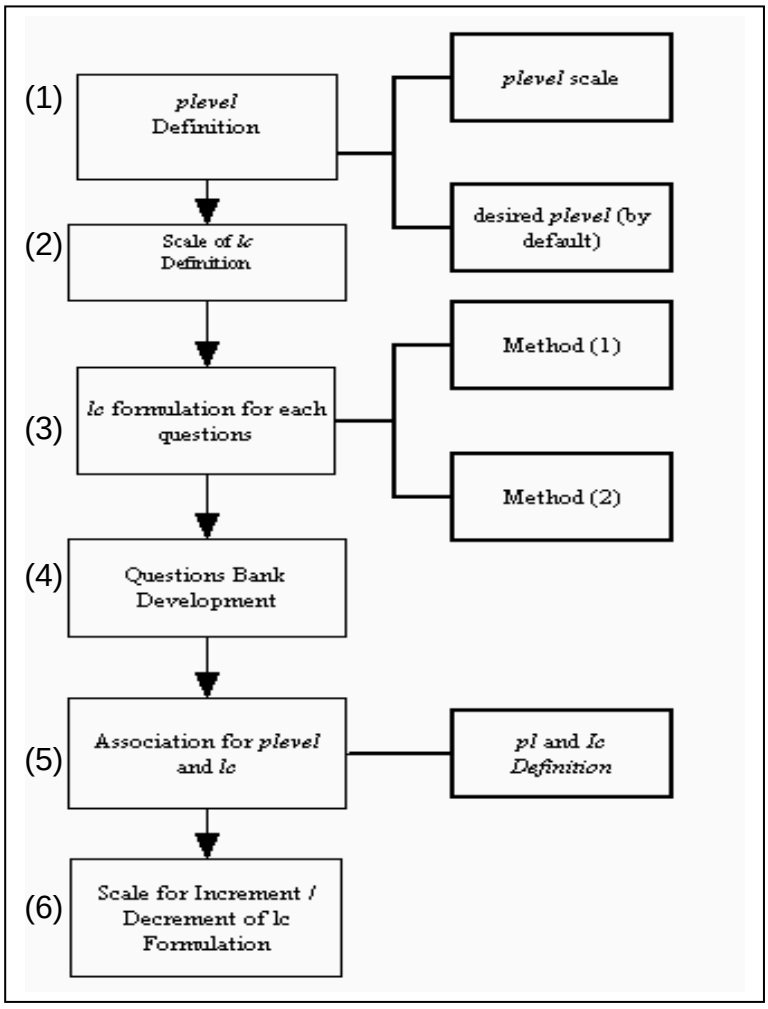


Figure 3.6: Generic Model for iCAP Computation.

3.3 Results and Benefits of iCAP Model in Blended Learning Environment

- **Progress Profile**
  Each learner's progress is captured and stored. The learners can always access the individual progress profile to identify their current state of performance versus their desired performance level. The lecturer can easily assess the learner's performance to identify each learner's performance and the progress of improvement or decline. Necessary action can be taken from this point.

- **Expandability and Flexibility**
  Expandability and flexibility means that this model is able to be expanded and adapted to a variety of requirements for lecturers. For instance:
  (1) The lecturer has the flexibility to determine the desired performance level and the level of challenge for each question.

(2) The lecturer is allowed to add, edit, and delete the questions in the question bank. The lecturer can also change the question difficulty level if necessary.
(3) iCAP can be applied to any content of teaching material.
(4) The lecturer can define the test which falls into the category of learners' tutorials or formal examinations.

- **Intelligent**
  Once the computation model is completed, the system developed based on the iCAP design is intelligent and able to generate many sets of test questions aligned to the individual learner's current performance level. The assessments and the learner's progress are captured and stored automatically.

3.4 Limitations of iCAP Model in Blended Learning Environment

- The questions model designed in iCAP has best fit with "Multiple Choices", "Fill in the Blank" or "True or False" type of questions. Essay or short comprehensive questions are difficult to be assessed unless another intelligent essay marking system is embedded with iCAP.

- The process of the model is tedious from the lecturer's perspective especially in stage (3). Although it is upfront effort for the lecturer at this stage, the learners can experience the benefits later.

- The definition of level of challenge can vary from one person to another. An assumption is made in iCAP – based on the majority'.

- Lecturer acceptance is not assured, with many educators doubting the ability of multiple-choice testing to assess higher order skills, and be a fair reflection of a student's knowledge. Many lecturers see multiple-choice as providing the students with the answer, it does not judge their knowledge (Davies, 2002).

4.0 Conclusion

This model is particularly useful for formative assessment where an iterative learning process is desired; learners can test themselves repeatedly on the same subject but with varied questions set to identify their current level of performance to the lecturer's expected performance. It plays a vital role from the lecturer's perspective because much attention is given towards individual learner's progress and the accessibility is wider and more effective.

Key advantages of the iCAP are the appreciation of individual learner's performance by educators and it acts as a motivation for learners to achieve their expected performance.

Future work for iCAP is to design an improved generation of assessments will be designed for computers making full use of essay, audio, video and

advanced graphics to enable complex questions and simulations. Such developments mean that ensuring quality will require a new and sophisticated range of measures from the level of the question to the level of educational management purposes (McKenna, 2000). The next wave of blended learning is ‘Education Unplugged’. This represents an evolution of blended learning that leverages the portability and utility of mobile and personal devices (Wagner, E., D., 2006). iCAP can be shifted from a web-based model aligned to the next generation of learning which is more personalised and customised based on the individual learner’s and educator’s needs.

References

Cohen, L., Manion, L., and Morrison, K. (2000). *Research Methods in Education*, London: RoutledgeFalmer.

Davies, P. (2002). *There's No Confidence in Multiple-Choice Testing…*, Proceedings for 6th CAA International Conference, 119-130.

Graham, C., R. (2006). 'Blended Learning Systems', in C., J., Bonk, C., R., Graham (eds), *The Handbook of Blended Learning: Global Perspectives, Local Designs.* CA: Pfeiffer, 5.

Inge, W. (1919). **'**The Training of The Reason', in A. C. Benson, (ed.). *Cambridge Essays on Education*, Cambridge, 15.

Jay, M. (2003). *The Education of John Dewey*. Columbia University Press.

Jones, N. (2006). 'E-College Wales, A Case Study of Blended Learning.' in C., J., Bonk, C., R., Graham (eds), *The Handbook of Blended Learning: Global Perspectives, Local Designs.* CA: Pfeiffer, 182-194.

McKenna, J. B. (2000). *Quality assurance of computer-assisted assessment: practical and strategic issues*. Journal of Quality Assurance in Education, 8(1) 24-31.

National Audit Office report on "Improving Student Achievement in English Higher Education" 18th January 2002
http://www.nao.org.uk/publications/nao_reports/01-02/0102486.pdf

Quality Assurance Agency (QAA) code of practice for assurance of academic quality and standards in higher education assessment on students May 2000 http://www.qaa.ac.uk/academicinfrastructure/codeOfPractice/section6/default.asp

Thorne, K. (2003). *Blended Learning - how to integrate online & traditional learning,* London: Kogan Page Limited.

Wagner, E., D. (2006) 'On Designing Interaction Experiences for The Net Generation of Blended Learning', in C., J., Bonk, C., R., Graham. (eds), *The Handbook of Blended Learning: Global Perspectives, Local Designs.* CA: Pfeiffer, 41-53.

Walker, J., C. (1999). 'Self-determinations as an Educational Aim', in Roger Marples (ed.) *The Aim of Education*, London: Routledge, 112-123.

Appendix A – Screen Shots of the System Implemented by iCAP model.

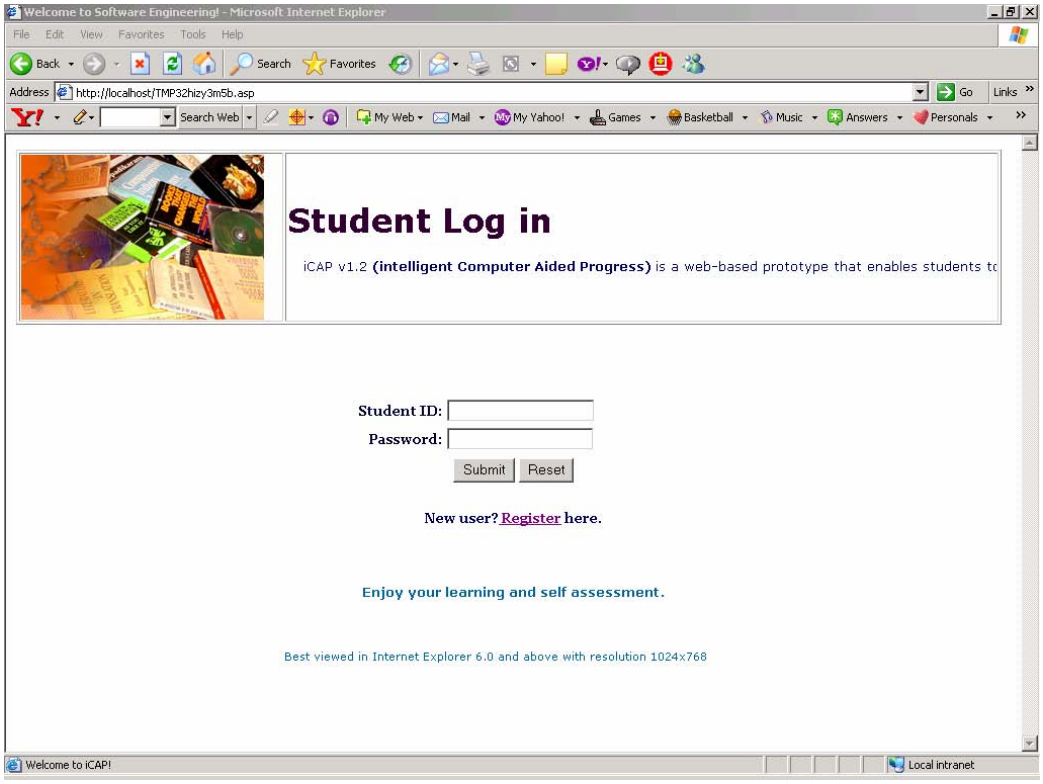


Figure 4.1: Learner's Log in Page

Figure 4.2: Learner's Main Page

Figure 4.3: Learner's Progress Profile

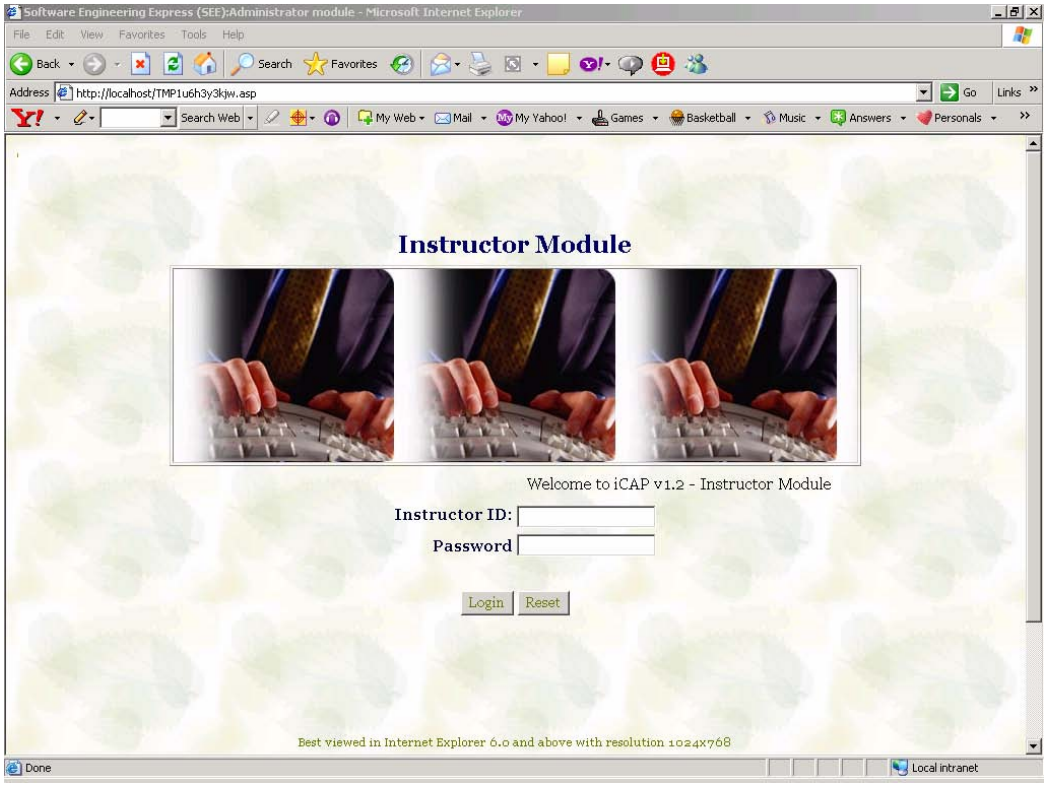


Figure 4.4: Instructor's Log In Page

Figure 4.5: Instructor's Main Page

IN-DEPTH CASE STUDIES OF STUDENTS' USE OF TECHNOLOGY TO SUPPORT ASSESSMENT

Gráinne Conole, Maarten de Laat, Jonathan Darby and Theresa Dillon

In-depth Case Studies of Students' Use of Technology to Support Assessment

Gráinne Conole, Maarten de Laat,
Jonathan Darby and Theresa Dillon
g.c.conole@open.ac.uk

A review of over eighty studies which purported to focus on students' experiences of e-learning highlighted some surprising results (Sharpe et al. 2005),[1] finding that few studies actually focused on the student experience. The JISC Learning Experience Project (LXP) is working with four of the HE Academy subject centres[2] to explore students' experiences of technology; with a particular interested in discipline difference in the use of technology for assessment purposes. The primary aim is to distil out subject discipline issues in using e-learning. This is being achieved by: collecting data on students' experiences of using technology to support learning activities, describing the students' personal background and learning context, and drawing out learner beliefs and e-learning strategies. After this initial situated exploration the focus will be turned to a wider set of issues involving learner's experience of both learning and technology and learner's thoughts and believes about their experiences.

Data collection includes an online survey, twenty in-depth case studies (including audio logs, interviews and observation) and focus groups. The research questions include:

- How do learners engage with and experience e-learning?
    - What is their perception of e-learning?
    - What do e-learners do when they are learning with technology?
- What strategies do e-learners use and what is effective?
- How does e-learning relate to and contribute to the whole learning experience?
- How do learners manage to fit e-learning around their traditional learning activities?

An important part of the study is to explore how students are using technologies to support their assessment activities; both in terms of creating assignments and undertaking online formative and summative assessment. We are interested in exploring the subject discipline differences in the types of assessment and the ways in which it is used. Evidence suggests that there

[1] Sharpe, R., Benfield, G., Lessner, E., & DeCicco, E. (2005) Final report: Scoping study for the pedagogy strand of the JISC learning programme. Unpublished internal report v.4.1 JISC.
[2] Medicine, Dentistry and Veterinary Medicine, Economics, Information and Computer Sciences and Languages and linguistics

are fundamental subject disciplines in the key characteristics of learning which impacts significantly on modes of assessments undertaken. For example a recent symposium highlighted the importance of communication in the Social Science, problem-based learning in Sciences and team work in Health Sciences.[3]

The paper will draw out the findings from the LXP study in relation to students' use of technologies for assessment. Initial analysis of early data from the study shows that students are conscious of both the benefits and disadvantages of e-assessment as the following quote illustrates:

> "*My experience is that it [e-assessment] certainly helps with formative assessment so that one can test oneself against different parts of the curriculum. The downsides include lack of personal feedback so that you don't necessarily know that what you are studying is what you should be studying.*"

The paper will draw out the key findings from the study in relation to e-assessment and use these as a basis for making recommendations for more effective e-assessment across different subject disciplines.

[3] HE Academy/JISC symposium – 9th February 2006, http://www.heacademy.ac.uk/eLDisciplines.htm

USE CASES AS A MEANS OF CAPTURING E-ASSESSMENT PRACTICES AND INDENTIFYING APPROPRIATE WEB SERVICES

Gráinne Conole, Isobel Falconer, Ann Jeffrey and Allison Littlejohn

Use Cases as a Means of Capturing E-assessment Practices and Identifying Appropriate Web Services

Gráinne Conole, Isobel Falconer,
Ann Jeffrey and Allison Littlejohn
g.c.conole@open.ac.uk

The JISC-funded LADIE project has produced a set of use cases of learning activities derived through a series of workshops with practitioners (*www.ladie.ac.uk*). From these an e-learning framework identifying the services needed to support learning activities has been produced.

The Learning Activity Reference Model (LARM) is part of the e-framework programme, to encourage people to design learning activities using appropriate technologies. A reference model such as the LARM provides a process for designing and implementing effective learning activities, from initial design, through requirements specification, to analysis of the technologies, specifications and standards necessary to meet those requirements. It identifies common requirements of reusable learning activities based in effective practice; and describes how these requirements can be met using existing and developing technologies, specifications and standards using a web services approach.

LADIE aimed is to provide a bridge between the plethora of learning activities which practitioners might wish to develop and identification and implementations of appropriate web services to support these. This presentation will focus on the assessment dimensions articulated in the use cases and how these are mapped in the LARM. It will critique the pedagogical aspects of e-assessment as highlighted in these use case, by attempting to draw out the relationship between particular pedagogical approaches, tasks undertaken by the students and associated assessment activities.

The presentation will give an overview of how the use cases were collected, demonstrating how the workshop material build on the DialogPlus Learning Activity taxonomy. It will go on to draw out the assessment dimensions evident in the use cases and show how they are mapped to particular web services. Finally, the presentation will give an overview of the LARM, described through three separate guides. Each guide is intended for a different audience:

- **Teachers / Practitioners:** the Pedagogy guide which has teaching and learning as its primary focus

- **Technologists / Implementers:** the Implementation Guide which describes how to configure learning activities from existing environments

- **Developers / Vendors:** the Services Guide which defines the reference model so that those creating new educational technology applications can ensure they can be used through the LARM.

It will focus in particular on the first of these - the Pedagogy Guide, which is designed for use by teaching practitioners who need to design and implement learning activities. It offers guidance on how to create a learning activity, on effective use of tools and resources in implementing activities, and a language and structure by which teaching practitioners and learning technologists might discuss the development and implementation of learning activities.

Contents

Draaijer S van Boxel P Summative Peer Assessment Using `Turnitin' and a Large Cohort of Students: A Case Study

Ellis E Greenhow M Hatt J Exportable Technologies: Mathml and SVG Objects for CAA and Web Content

Farrell G A Comparison of an Innovative Web-based Assessment Tool Utilizing Confidence Measurement to the Traditional Multiple Choice, Short Answer and Problem Solving Questions

Harrison G Gray J A Computer-assisted Test for Accessible Computer-assisted Assessment

Head S Ogden C Development of a Student-searchable Database of Veterinary MCQ's with Educational Feedback for Independent Learning

Hermet M Szpakowicz S Symbolic Assessment of Free-text Answers in a Second Language Tutoring System

Johnson R Johnson S Generalise not Specialise: Design Implications for a National Assessment Bank

Khan S Maple TA: A Springboard for Web Based Testing and Assessment

MacKenzie D Stanwell M QuickTrI and Intelligent Shell System (ISS) from Innovation 4 Learning. Building on Practical Experience

Mann H Glasfurd-Brown G Learning from Assessment: Evaluating the Benefits of DALI (Diagnostic Assessment Learning Interface)

Martin H Formative Assessment Using CAA: An Early Exploration of the SLIM Pilot Project

McAlpine M van der Zanden L Itembanking Infrastructure: A Proposal for a Decoupled Architecture

McGee Wood M Jones C Sargeant J Reed P Light-weight Clustering Techniques for Short Text Answers in Human Computer Collaborative (HCC) CAA

McGill L Overview of JISC Assessment activities

Mogey N Sarab G Exams, Essays and Tablet Computers – Trying To Make the Pill More Palatable

Moody J Swift J Development of Web Browsing Techniques to Capture Responses in the Context of English Language Skills Assessment

Nicol Dr D Assessment for Learner Self Regulation: Enhancing the First Year Experience Using Learning Technologies

Osborne C Winkley J Developments in On-Screen Assessment Design for Examinations

Pickard P Assessment to Improve Self Regulated Learning

Qudrat-Ullah H System Dynamics Based Learning Environments: A Technology for Decision Support and Assessment

Sangwin CJ Mathematical Question Spaces

Sclater N Butcher P Thomas P Jordan S Moodle: Enhancing the Assessment Capabilities of the Leading Open Source Virtual Learning Environment

Shepherd E Innovations in E-Assessment

Sim G Read JC Holifield P Evaluating the User Experience in CAA Environments: What Affects User Satisfaction?

Trinder JJ Magill J Roy S A Call to Arms for Handheld Devices

Tulloch I CATS - Constructing Assessments using Tools and Services

Warburton B Quick win or Slow burn? Modelling UK HE CAA uptake

Whaley H Walker D Development of a Web-based Groupwork Assessment Tool

Wheadon C He Q An Investigation of the Response Time for Maths Items in A Computer Adaptive Test

Whitehouse G Intelligent Paper, Pens and Ink

Whitelock D Brasher A Developing a Roadmap for e-Assessment: Which Way Now?

Whitelock D Mackenzie D Whitehouse C Ruedel C Rae S Identifying Innovative and Effective Practice in e-Assessment: Findings from Seventeen UK Case Studies

Wills G Davis H Chennupati S Gilbert L Howard Y Jeyes S Millard D Sherratt R Willingham G R2Q2: Rendering and Reponses Processing for QTIv2 Question Types

Wynne L Lopes S Implementing Large Scale Assessment Programmes

QUESTIONBUDDY – A COLLABORATIVE QUESTION SEARCH AND PLAY PORTAL

Will M Davies and Hugh C Davis

QuestionBuddy – A Collaborative Question Search and Play Portal

Will M Davies and Hugh C Davis
Learning Technologies Group
ECS
University of Southampton
Southampton
UK
wmd04r@ecs.soton.ac.uk

Abstract

Generally itembanks are inaccessible to students. Current use of itembanks focus on the teacher as having responsibility to organise questions (place them in pools, associate them with course content) and make them available/deliver them to students. This limits students to the teachers perspective and to the questions that the teacher has made available. As the practice of itembanking increases it may be appropriate to encourage students to use questions from pools not directly prepared by their teacher. A mechanism for searching across itembanks and sharing recommendations with peers would be of help in facilitating this. We describe QuestionBuddy, a collaborative filter based question portal for students, built to study student usage of, and attitudes to, such a system.

Introduction

We introduce QuestionBuddy, our self assessment website for students of electronic and electrical engineering. The site allows students to search for questions from the (E3AN) itembank and gives feedback to their answers. Having attempted a question the student is then asked to rate it, on relevance to their current study. By comparing a student's rating profile with those of other users, recommendations for further study questions can be made. This is done by selecting additional items rated highly by users with a similar rating profile.

The reason for this work is to investigate ways of enabling users to find itembank content for their needs. It is assumed that searching across the item metadata alone will not always be able to offer a complete solution to satisfy users' search requirements. Factors contributing to this include varied granularity of metadata and possibly the users incomplete knowledge of the domain they are searching. It is intended that that this work will be able to contribute in the area of engaging students and assisting them in seeking feedback. The QuestionBuddy self-assessment process aims to help students to making informed choices when directing their study efforts. By using the system regularly, students will be able to get timely feedback on the

effectiveness of their study. It is anticipated that lessons learnt from QuestionBuddy will be applicable to other sets of itembank users, such as teachers, that compile assessments. A call for an improvement in user/itembank interfaces can be seen in the Itembanks Infrastructure Study [IBIS], (Cross 2004).

QuestionBuddy has been built using the content from the E3AN itembank in combination with the APIS rendering engine, available at (APIS), for questions in IMS Question and Test Interoperability [QTI] format. To complement these, a custom webservice search interface has been added to E3AN and a collaborative filter has been constructed to make item recommendations to users of the site. To enable the APIS service to handle question rendering and response processing the original E3AN questions were converted to QTIv2 using the (PyAssess) conversion tool.

The Problem

Hidden Content

As the size and availability of learning object repositories and itembanks increase teachers are about to be swamped by yet another source of learning resources. Developing tools and techniques for finding and managing these resources is crucial. (Anderson, Ball et al. 2003) discuss the issues raised in searching for ever smaller learning objects with increasingly fine grained descriptions. (Lemire, Boley et al. 2005) identify problems in searching for learning objects over subjective metadata such as the IEEE Learning Object Metadata classification ‘semantic density’.

Search systems in existing itembanks rely on author/librarian created metadata, pools of questions created by teachers and crude plain text searches. To return questions appropriate to the student’s needs these techniques require significant input from the author/librarian/teacher in classifying the questions. When a question is used outside the context which it was created for it is likely that its description will need to be reconsidered. This is a significant problem for an itembank that is intended to be shared among a large number of institutions. It seems reasonable to consider search and retrieval issues relating to a single objective question as similar to those associated with a small learning object.

A Possible Solution

It is possible to gain knowledge about an item from its previous usage. In traditional models of summative assessment this usually means recording student scores and carrying out analysis of these scores. This can be used to identify questions that unfairly discriminate against certain students and also to identify discrepancies between the taught curriculum and the subject assessed. Having identified unfair questions it is then possible to remove them from future use. This analysis relies on results from a significant number of students. Rather than asking one expert whether, in their opinion, a question is biased, statistics make it possible to examine the results of a large number of students.

Extending this analysis to a formative assessment environment used by students from multiple institutions, with varying curricula, at different stages of their courses, appears fraught with statistical problems. The need for investigation in this area is stated in the IBIS report by (McAlpine and Cross 2004).

"As the analysis of student data is generally for summative purposes, a closer look must be taken at this use to facilitate formative use and empower students and their learning. Some of the key ways that this can be done is through helping students to make the correct choices in their learning by providing them with data which can assist them become more responsive and self-aware learners"

Collaborative filtering provides a way of making comparisons between similar users. In its simplest terms collaborative filtering ignores the all properties of an item except for the identities of the users that have interacted with it. The commonality between two items is measured by the intersection of the sets of people that have used them. For a class of 100 students on a course, they are no longer 100 individuals struggling in a sea of questions to find revision material, they can be empowered with the results of each other's efforts.Rather than relying purely on the use of protected term classifications and placing the sole burden of describing a question accurately with the author/librarian, we aim to augment an item's description with some notion of the context in which it is used.

Collaborative Filtering Overview

Successful collaborative filter systems include those used at (Amazon.com; Last.fm; MovieLens). Simplistically, a collaborative filter works by comparing two user's ratings of some material and calculating the similarity or distance between these users. If two users have a high degree of similarity then it is assumed that they will appreciate recommendations of items that they have not rated, but have been rated highly by the other. For an in depth review of collaborative filtering systems and techniques see (Adomavicius and Tuzhilin 2005), many of the systems they discuss take a hybrid approach of combining collaborative and content-based recommendations.

One collaborative filter that is likely to be familiar to many people is that used by Amazon.com. By making comparisons between different users' purchase and rating profiles Amazon is able to suggest items for purchase. The usefulness of these recommendations varies, one reason for this is that the system does not record the context in which a purchase is made. A good example of this is that by buying gifts for several very different people a customer can end up getting recommendations from several conflicting stereotypes. Amazon has recognised this and now includes a link with each item, 'why was this recommended to me' that allows users to remove items from their rating profile. (Linden, Smith et al. 2003) describes the specific filtering algorithm used by Amazon.

Collaborative filters are well established technology but they have not, until now, been used for question material. In the education domain, (Downes, Fournier et al. 2004) discuss Sifter, an experimental learning object recommender developed in Canada. The filtering system behind this, RACOFI, is now being used to power (InDiscover) a music recommender system. Sifter asked users to rate content along up to 15 dimensions including 'level of interaction' to 'ability to motivate'. One of RACOFI's strengths is its ability to filter efficiently over a large number of dimensions. The intended Sifter user group were developers responsible for assembling learning objects to build coherent courses.

There is a trade-off when creating a collaborative filter, between obtaining a sufficiently rich user model, and overloading the user by asking them too many questions about themselves and the content they are using, (Swearingen and Sinha 2002). In the context of Sifter, asking developers using a learning object to apply ratings in 15 dimensions seems acceptable. Asking for a similar level of detail from a student taking a two minute multiple choice question creates a burden that the student is unlikely to tolerate.

QuestionBuddy – The User Experience

Students are expected to come to QuestionBuddy having already studied a subject but wanting to confirm their understanding. After logging in they are presented with a personal summary page, shown in Figure 1, this presents some recommended questions. These recommendations are created both by analysing previous subject interests and also by the collaborative filter.

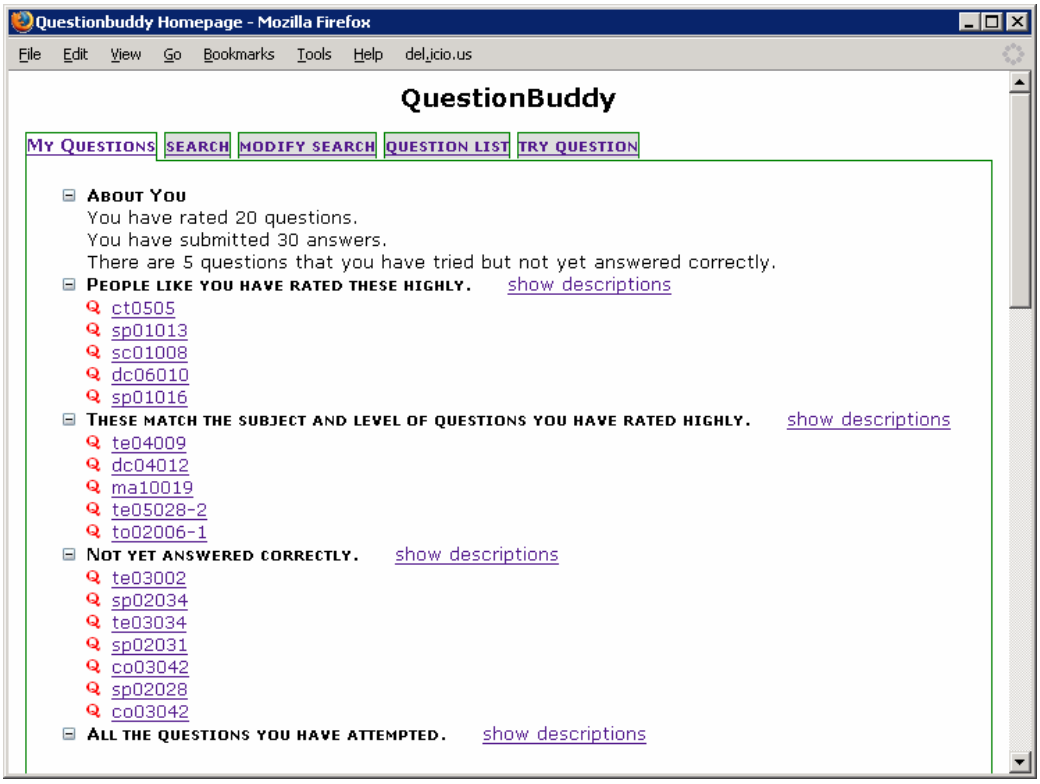


Figure 1

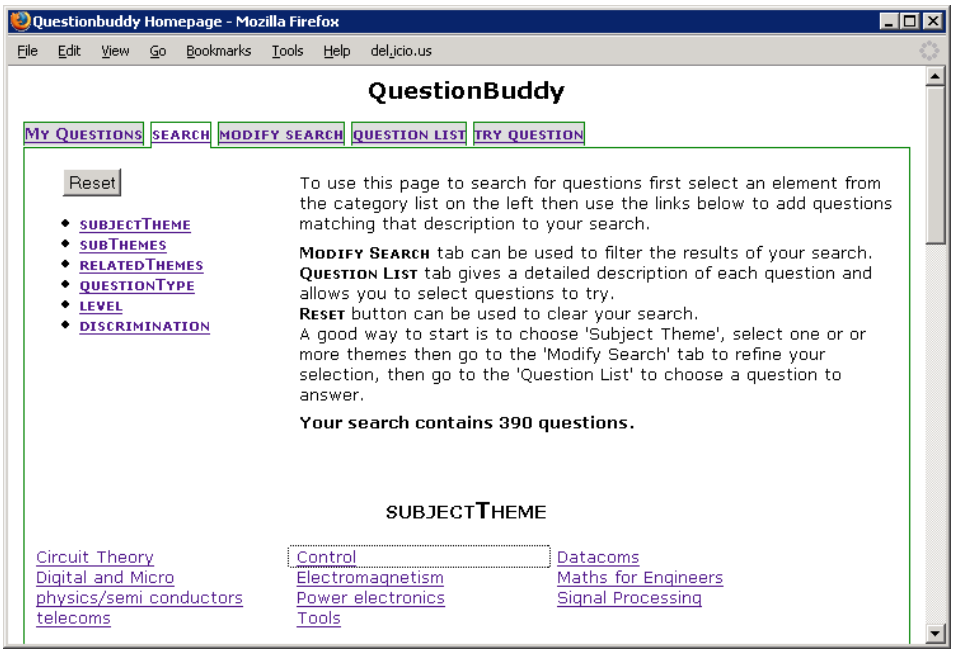


Figure 2

For a student new to the system this page will not be able to make recommendations, they will need to use the search page, Figure 2. The *search* page presents closed lists of categories from which users can select the questions that interest them. The number of hits in their search is updated and displayed as the scope of the search is increased.

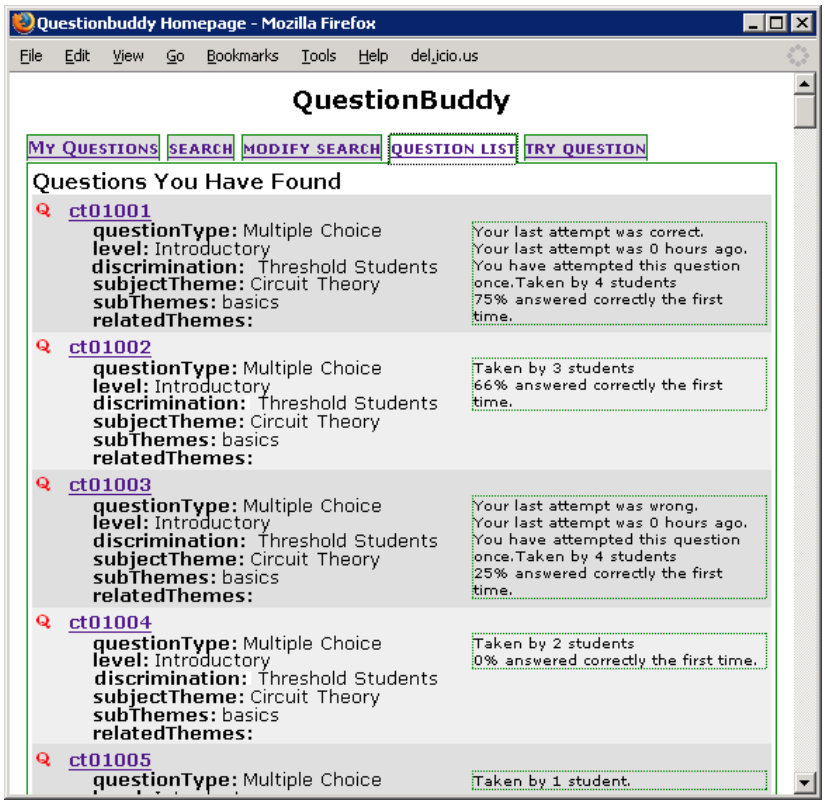


Figure 3

If desired the *modify search*, not shown, page can then be used to filter the search for example by only including questions that are multiple choice. The student may also choose to restrict the difficulty, discrimination or sub-theme of the results to reduce the number of returns to a manageable number.

Navigating to the *question list* page, Figure 3, displays the results of the students search. Each item is described using the available metadata and also some statistics concerning its previous use. This description is one of the areas of the system that needs further investigation. Important design questions are, what, of the information presented, is useful, and, what other information could be shown to help users.

Selecting a question from the list takes the student to the *try question* page, **Figure 4**, where the question is displayed and the student can submit an answer. If the question type chosen is supported by the QTI renderer it will examine the students answer and give them feedback. The ratings panel is provided for students to rate the question for relevance. They are required to submit a rating before the system will allow them to navigate away from this page. At present the question answer process requires users to navigate back and forth between the *question list* and *try question* pages, it is recognised that this impacts on the usability of the system. Consideration is being given to allowing each of the questions in the *question list* to be displayed inline without forcing users to navigate between panes.

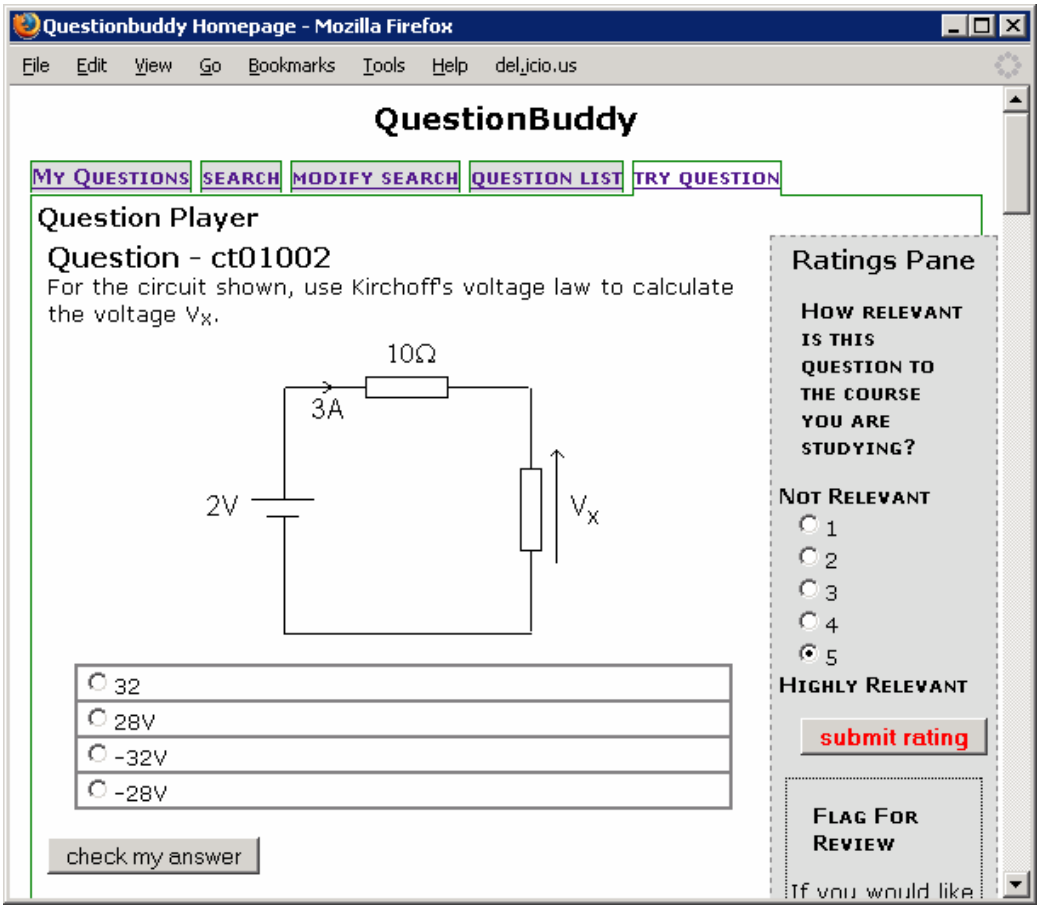


Figure 4

System Architecture

QuestionBuddy is implemented by aggregating several webservices. These services are: the APIS QTIv2 renderer and home grown services for maintaining user profiles and itembank searching.

Itembank Search Service

The search service is implemented on top of the E3AN itembank of electronic and electrical engineering questions. The interface to the search service provides four methods:

- getSearchTerms()
- getSearchTermValues(String searchTerm)
- search(String query)
- completeTentativeSearch(String searchIdentifier)

In addition four objects SearchResult, SearchTerm, SearchTermValue and Item are required by the interface. The service is designed to return lists of searchable terms rather than expecting users to guess how the content has been categorised. Whilst this adds extra complexity to the user interface, it should simplify the construction of sophisticated queries. The search service protocol places no restriction on the way questions are categorised so it would be possible to aggregate results from several itembanks if this is desired. The protocol has been kept deliberately simple to allow compliant services to be created for existing itembank systems. A version of this webservice search interface has also been implemented for the TOIA itembank. No changes were necessary to the client to allow this interface to work successfully.

QTI Rendering and Response

The APIS rendering and response service as downloaded from sourceforge required a small number of changes to the code to generate correct XML and to handle the QTI expression *match*. We look forward to integrating the R2Q2 QTI webservice renderer that is being funded under JISC toolkit development. This should extend the range of questions types that QuestionBuddy is able to play.

User Profiling and Collaborative Filter

The user profile service was developed independently from the itembank service. This was done to ensure that any developments made to the service were independent of the itembank used by the system. This service will work successfully with multiple itembanks providing the item identifiers are unique throughout the system.

Lessons Learnt

Trying to create meaningful descriptions of items to display in a list for students is not easy. A similar problem would be asking someone with no knowledge of science fiction to choose a science fiction book as a gift. With little knowledge of the sub-classification of the genre much of the information they could be shown about the book will be meaningless. The solution chosen for QuestionBuddy works best when users understand the specific educational language used in the metadata. This display is augmented with statistics of previous question usage. The collaborative filter should help to compensate for less than ideal question descriptions by ensuring that a greater proportion of the questions offered are relevant.

The system contains more information about each question than it is useful to present to the student when helping them to choose questions to attempt. In part this is caused by the specific/specialist nature of some of the metadata. For example, E3AN contains a description for cognitive level, indicating what level of skills the question assesses. This information is likely to be helpful to a teacher compiling an assessment but is probably not helpful for the target student audience. As a result of this more than 50% of the data about a question provided by the search service was discarded.

The decoupling of the search parameters from the user interface complicated the user interface design. This feature is important to enable the interface to work with different itembanks. Knowing how many categories existed and how many possible values they could take, would allow for a more intuitive interface design.

It is possible to calculate an average rating for each item, but as the rating depends on the context of the student this would ignore the fact that different students will be studying different, if subtly, courses. As a consequence a definite decision was taken not to display the average rating of an item.

Future Work

Once the system is in regular use it will be possible to look for trends in rating. It may be possible to use these to learn more about the content and to tune the recommendation system. One way of doing this is by analysing item ratings in conjunction with the search criteria used to find them. For example, if users searching for questions on ‘circuit theory’ always rate question X 1/5, either this question is not about circuit theory or, it is simply not a very useful question.

After calculating the discrimination of each question it may become desirable to filter out questions with low discrimination. This would ideally allow students to take fewer questions to get an accurate assessment of their ability. Because of the formative nature of the system, this is problematic as hopefully the students’ ability is improving from one session to the next.

In a much the same way as Amazon allows customers to remove certain purchases from their recommendation profile, it might be helpful to allow students to specify constraints on their recommendations. For example a student that has taken 90% of the questions on electromagnetism is likely to get recommendations for the other 10%. The student may feel they have studied this area sufficiently and wish to exclude these questions. It should be possible to make this decision automatically by examining the students performance in previous questions. This is related to a more general issue that the recommendation system should be transparent. The system should be capable of displaying to the user how their recommendations are generated and wherever feasible they should be able to adjust the parameters controlling what is offered to them.

Social bookmarking, the act of creating personal tags for collections of resources is currently a popular way of allowing users to describe things for their own and others use. For a good introduction to tagging see (Wikipedia). Allowing users to create a folksonomy of an itembank may create harvestable information about the question held. In conventional itembanking terminology, this is very similar to pool creation. The ability to create pools and share pool identifiers with other users would support other use cases for the system. This type of feature needs to be examined carefully as malicious users could create deliberately disparate collections that might poison the system for others.

Conclusion

QuestionBuddy is ready to offer students a novel way of self-testing. By analysing the use of the system in combination with the existing item metadata we anticipate being able to augment the user experience. It will be possible to utilise the usage data recorded about each item to increase the value of the item in the future.

References

Adomavicius, G. and A. Tuzhilin (2005). "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions." Knowledge and Data Engineering, IEEE Transactions on 17(6): 749.

Amazon.com, http://www.amazon.com, last accessed: 01/04/2006.

Anderson, M., M. Ball, et al. (2003). RACOFI: A Rule-Applying Collaborative Filtering System. COLA'03, Halifax, Canada, IEEE WIC.

APIS, http://apis.sourceforge.net/, last accessed: 01/04/2006.

Cross, R. (2004). Item Banks Infrastructure Study. N. Sclater, JISC**:** 87-100.

Downes, S., H. Fournier, et al. (2004). Projecting Quality. MADLaT, Winnipeg,Manitoba.

E3AN, http://www.e3an.ac.uk/, last accessed: 01/05/2006.

InDiscover, http://www.indiscover.net, last accessed: 01/04/2006.

Last.fm, http://www.last.fm, last accessed: 1/04/2006.

Lemire, D., H. Boley, et al. (2005). "Collaborative Filtering and Inference Rules for Context-Aware Learning Object Recommendation." International Journal of Interactive Technology and Smart Education, 2(3).

Linden, G., B. Smith, et al. (2003). "Amazon.com recommendations: item-to-item collaborative filtering." IEEE Internet Computing 7(1): 76-80.

McAlpine, M. and R. Cross (2004). Item Banks Infrastructure Study. N. Sclater, JISC**:** 77-85.

MovieLens, http://movielens.umn.edu/main, last accessed: 01/04/2006.

PyAssess, http://pyassess.ucles.cam.ac.uk/, last accessed: 01/05/2006.

Swearingen, K. and R. Sinha (2002). Interaction Design for Recommender Systems. Designing Interactive Systems (DIS2002), London, ACM.

Wikipedia, http://en.wikipedia.org/wiki/Folksonomy, last accessed: 01/04/2006.

THE MOBILE WIRELESS CLASSROOM: POCKET PC'S IN HIGHER EDUCATION

Christopher J. Dechter

The Mobile Wireless Classroom: Pocket PCs in Higher Education

Christopher J. Dechter
Teaching & Learning Center
Eastern Washington University
106 PAT
Cheney
WA 99004
USA
cdechter@ewu.edu

Abstract

Since 2003, EWU has been using the Mobile Wireless Classroom, a pilot project consisting of a self-contained portable set of 30 Pocket PCs for electronic assessments. Pocket PCs were selected as an alternative to laptop computers as they maintain much of the same capability but at a fraction of the cost and are much less invasive in general classroom use. Students ranging from first-quarter freshmen in English Composition to last-quarter seniors in Developmental Psychology use the Pocket PCs to respond to lectures, watch videos of laboratory procedures, and to submit writing samples for peer review. Professors using the mobile assessments can quickly gather feedback from students and pinpoint areas where further review is needed.

An Idea is Born

In February 2003, Ian Siemer (a colleague in my office) and I were thinking of ways to make computers more accessible to students in classrooms here at Eastern Washington University. We wanted an alternative to laptops, not only because of cost, but because laptops often create a barrier between instructor and student and can become a distraction in the classroom. We wanted a system with similar functionality to a laptop-equipped classroom, but with less intrusive technology and at a lower cost.

We then began investigating the possibility of creating a classroom set of Pocket PCs as an alternative to a much more costly set of desktop or laptop computers. At the time, EWU already had several sets of wireless-capable laptops, but they were assigned to a single classroom, generally required installed infrastructure (wireless access points and mounted projector), and were cost-prohibitive for many colleges and departments on campus.

Based on my personal experience using Pocket PCs, and Ian's insistence that students would be excited to use them in class, we posited that a set of Pocket PCs and related accessories designed to travel with them from

building to building and classroom to classroom would not only be an inexpensive alternative to laptops, but would be an attractive option for many departments and instructors on campus. Thus the Mobile Wireless Classroom was born.

Putting Together the Pieces

The Mobile Wireless Classroom (MWC) is a transportable, self-contained classroom set of handheld computers and a centralized server for electronic polling, quizzing, testing, assessment, and streaming audio & video, connected via an ad hoc wireless network. In a nutshell, the MWC is a set of 30 Pocket PCs stored in a custom built cart (complete with charging cradles, laptop server, projector, and wireless access points) that can be moved to any classroom on campus and used by any instructor in any subject.

For hardware, we chose the Dell Axim X5 (running Windows Mobile 2003) for student use and matched that with a Dell Latitude D600 (running Windows Server 2003) acting as both instructor laptop and server. We mounted 30 cradles and 30 spare batteries in a standard (read: ugly) wheeled cart, along with two Apple Airport wireless access points (Snow models, modified with external antennas), a BenQ PB7200 projector, and pre-wired everything to minimize setup time. Instructors have only to move the cart into their classroom, plug in a power cable to the wall, an optional network cable to the campus network, and the MWC is ready to go in about 90 seconds.

The network cable is optional because many applications of the MWC do not require an outside network connection, and simply not connecting to the campus network keeps students on task and away from Hotmail and eBay. Aside from network access, we did not otherwise limit the Pocket PCs because as a pilot project, we did not want to discount any possibility.

For software, we chose ‘QuestionMark Perception 3’ to drive online assessments and tests, and ‘TurningTechnologies TurningPoint vPad 2003’ for interactive lectures and quizzing and polling. We use ‘Sprite Clone’ for maintenance and cloning of the Pocket PCs, saving hours of configuration. Most navigation is done via an internal Web site and Pocket Internet Explorer, so development is all in HTML and can be easily updated.

We initially intended the MWC as a tool for classroom polling and basic quizzing, and even wrote a custom application to do just that, but after discussion with interested instructors realized not only that it had to do more, but that we were just touching the surface of the capabilities of wireless Pocket PCs in a classroom. We've since added on-demand streaming audio and video, more advanced web-based assessments, interactive lecture response, and are investigating in-class instant messaging.

The first public demonstration of the MWC was in May 2003, and we received many inquiries from instructors interested in participating in the pilot. We selected instructors from varying disciplines who expressed a diversity of ideas on how to use the MWC and the Pocket PCs in their classes. During the

summer of 2003, we met several times to discuss implementation, logistics, assessment, and pedagogy, as well as the technology.

Rollout

In Fall 2003, the Mobile Wireless Classroom rolled out (literally) to Microbiology classes, and got very positive reactions from the students and the instructor. Students individually reviewed on-demand streaming videos of laboratory procedures from a library of 30 videos. The instructor then followed up with questions about what they just saw, and could give students instant feedback.

Despite initial network problems (streaming 30 different videos to 30 Pocket PCs simultaneously via 802.11b presents many of its own problems), the students used the on-demand videos for several weeks during labs with few issues. During heavy network traffic, many of the Pocket PCs would not reliably stream videos. The videos, encoded at 200kbps, would seemingly choke the access points when more than 15 Pocket PCs were connected. Through trial and error, we found we got the best and most reliable performance using older model Apple AirPorts. We never did identify the exact issue beyond being able to replicate it with almost any brand or model access point, but the problem with the network was something of a blessing in disguise, as having to debug the most technically difficult project first provided us with an incredibly robust network configuration; there have been no problems with it in the 18 months since.

Expanding the Options

During early 2004, we expanded the use of the MWC to include in-class quizzing and polling via QuestionMark Perception, and dealt with issues arising from sharing the MWC between two instructors using it in different buildings on the same day. In Filmic Arts-Directing and Producing, the class viewed student films and then gave anonymous feedback, electronically and instantly. The student filmmakers then responded directly to critiques of their projects and even created questions specific to their own work. The student feedback was more honest than paper-based submissions because of the anonymity afforded by the Pocket PCs. This electronic feedback system replaced a paper-based system where feedback was not reviewed until days after the actual class, thus diminishing its usefulness.

In Electrocardiography Interpretation, students reviewed electrocardiograms in class, interpreted the content, and responded via electronic polls and quizzes. Logistical and scheduling issues aside (and an incident in which the cart was not plugged in over a weekend!), there were few problems.

In Fall 2004, students in English Composition used TurningPoint vPad and the MWC to participate in interactive lectures on grammar and mechanics and to submit their own writing samples for peer review and to discuss them in groups or as a class. While the instructor's original plans included only simple

classroom polling and lecture response via vPad, the addition of essays and open-ended questions, and student-submitted writing samples were very useful and popular with both the instructor and students.

In Spring 2005, students used the MWC, TurningPoint vPad, and QuestionMark Perception in Developmental Psychology to answer questions embedded throughout lecture presentations, and to receive immediate feedback about their responses in order to identify areas in which they need review. This helps instructors manage a dynamic lecture. Students also work in groups to compare individual and group responses. This continued throughout the 2005-2006 academic year.

Pocket PCs are not without their caveats (screen size, resolution, application availability), but in five university classes ranging from first quarter freshmen in English Composition to last quarter seniors in Developmental Psychology, Pocket PCs have proven to be valuable in classroom instruction. Screen size and the stylus input that are the biggest areas of concern to instructors and students, but as long as the classroom activities are designed to accommodate the lower resolution screens and input capabilities of Pocket PCs, these limitations have never caused problems. Students take to stylus input very quickly, and screen size primarily makes web access more cumbersome. Pocket PCs with VGA screens and built-in keyboards would all but eliminate these issues.

A Definite Success, with Promise for the Future

We're now approaching the end of our second year in a pilot program to determine the feasibility of Pocket PCs as a classroom enhancement and as an alternative to a dedicated computer classroom. The results have been very interesting. Our original idea was to see if wireless Pocket PCs could serve as a lower-cost and less invasive alternative to a dedicated computer classroom: in that I feel we've been very successful.

What future do Pocket PCs have in higher education? A very bright one if handled correctly. We found that if instructors treat them as direct laptop replacements, Pocket PCs often fail to impress. But if instructors look at Pocket PCs as a different tool for students to use in-class electronic assessments, audio, video, polling, and quizzing, then Pocket PCs work quite well. Pocket PCs can be an attractive option to many departments, and are very capable for a variety of uses. The MWC prototype system totaled less than $15,000 including development and personnel costs, while a comparable classroom set of laptops costs about $30,000 to $45,000.

We'd like to expand the MWC to include larger classes of 60-100 students, create additional sets for satellite campuses, and take the Pocket PCs off campus for use in the field in courses in the environmental sciences, forensics, marketing, social work, and business. The response from students and instructors thus far has been very positive, and as advances in handheld computing continue, so shall the possibilities for their use in higher education.

ONLINE COURSEWORK SUBMISSION FROM PILOT TO UNIVERSITY-WIDE IMPLEMENTATION: RATIONALE, CHALLENGES AND FURTHER DEVELOPMENT

Andrew Downton, Guinevere Glasfurd-Brown and Rob Mossop

Online Coursework Submission from Pilot to University-wide Implementation: Rationale, Challenges and Further Development

Andrew Downton, Guinevere Glasfurd-Brown and Rob Mossop
University of Essex
01206 872026
acd@essex.ac.uk
guin@essex.ac.uk
rwmoss@essex.ac.uk

Overview of the OCS Project

The paper outlines the development of a University-wide Online Coursework Submission system (OCS), which was funded by the University of Essex Teaching and Learning Innovation Fund in 2004-05, before being rolled-out across the University in 2005-06. The OCS project was informed by much smaller systems already running in three departments and a University-wide survey of departmental coursework submission requirements.

The project sought to establish a single system that would meet the needs of staff in departments, would facilitate coursework administration, management and quality assurance, and help to address rising workloads associated with these processes. The OCS system was also designed to support submission of coursework to the JISC Plagiarism Detection Service, now Turnitin UK.

The introduction of OCS coincided with the university's adoption of zero tolerance marking (ZTM) and one of the central reasons for adoption by departments has been to assist with the management and arbitration of the new ZTM policy.

The general drive behind the adoption of a university wide electronic submission system was not what might ordinarily motivate such a project, namely the need for supporting widespread electronic marking, indeed one of the key reasons uptake of OCS has been so rapid is precisely because the system does not force departments or individual members of staff to adopt electronic marking. Rather the central motivating factors lay with easing the administrative burden of ZTM and providing easier routes for departments to deliver on quality assurance, submission monitoring, facilitating JISC plagiarism checking as well as supporting staff who do wish to either receive work electronically (such as in code-based assignments in the Electronic Systems Engineering (ESE) and Computer Science (CS) departments) or wish to mark, in some form, electronically. The area of online marking is certainly one in which the University is interested, but it is likely that a pilot would be undertaken in the first instance.

Rationale

The use and uptake of VLEs within higher education is variable[i]; VLE versions also vary in terms of functionality. At the University of Essex there is no requirement on staff to make use of the VLE, WebCT, and the version does not interface fully with University Management Information Systems (MIS) and Student Records Databases (SRDB) . Whilst VLEs have online submission tools for coursework submission, the onus is on staff to enable submission for any courses that they run within the VLE. The experience for both staff and students is therefore quite uneven. It was clear that there was need to implement a complementary system tailored to University requirements, which could be taken up by larger numbers of staff.

Usability was a key aim of the OCS system. The project team sought to establish a system, which would not require any set-up by staff, (unlike a VLE) The OCS was embedded within the University Student Portal to facilitate access by students.

In 2005-06 the functionality and interface of the OCS system were improved to accommodate two key issues highlighted by the pilot process:

Scalability: the pilot system was extremely labour intensive from a systems management point of view. Each course and assignment together with student upload directories and permissions had to be setup manually, which meant that when broadened out to include, potentially, all departments and hundreds of courses scalability became a serious issue.

Electronic / hard-copy disparity: outside of a context in which the work would be viewed and/or assessed in electronic form, students were still required to submit a hard-copy version of their work alongside their electronic submission. Departments raised valid concerns at the pilot stage that there is potential for students to submit one version of their work electronically and another in hard-copy, thus circumventing the effectiveness of electronic plagiarism checking. Additionally, one department involved in the pilot did not have simultaneous hard-copy and electronic deadlines as work was submitted in class rather than to a central office meaning that students with classes later in the week could attempt to gain additional time by submitting incomplete work online and continuing to work after some of their fellow students had submitted their work at an earlier class, which is clearly unfair. In general terms it became clear that some kind of mechanism was required to ensure, as far as possible, that work presented in hard-copy form would match that submitted online in electronic form.

These two issues, together with the requirement that adoption of electronic submission should not equate to an adoption of electronic marking, which would have drastically reduced uptake of the final system, meant that the delivery of electronic submission via a VLE (in our case webCT) was not feasible.

To begin with, VLE integration with our MIS student records and courses databases is minimal, to the extent that there would have been little difference in terms of scalability between operating electronic submission via the pilot system and via a VLE – in fact the former would have been the preferred option had it been a choice between the two as there was less work involved

in setting up student directories for an assignment under the pilot system than there would be under the VLE. Given that the requirement was for a system that required minimal additional input, and that any setup workload could be handled by departments' administrative sections, an in-house system appeared to be the only way to achieve the aims of the project.

The in-house solution becomes more pressing when considering the additional problem of version disparity between electronic and hard-copy submissions. VLE's do not commonly consider the issue of this kind of disparity. To deal with the possible disparity between hard-copy and electronic copy a facility known as 'watermarking' was developed. After submitting their work (the system supports a variety of formats) to the system, students select the 'watermark' option. This requests a special copy of their work to be produced, which is delivered in PDF format to the student via the upload page for that assignment. The student then downloads the watermarked file and prints it off for hard-copy submission. The 'watermark' is a string of information that the student cannot derive independently of requesting a watermarked copy through the system and appears on every page of the document. It is the printed version of this file that students must submit to their department, who can check the authenticity of any particular watermark by comparing it to the reports produced by the OCS system.

Together, these factors formed the basis of the decision to pursue an in-house rather than a VLE-based solution to the university's electronic submission requirements. The final version of this solution was implemented utilising ASP.Net web forms for the staff and student front ends and .Net windows services that provide the back-end functionality responsible for the management of the watermarking system and production of assignment zip files and reports.

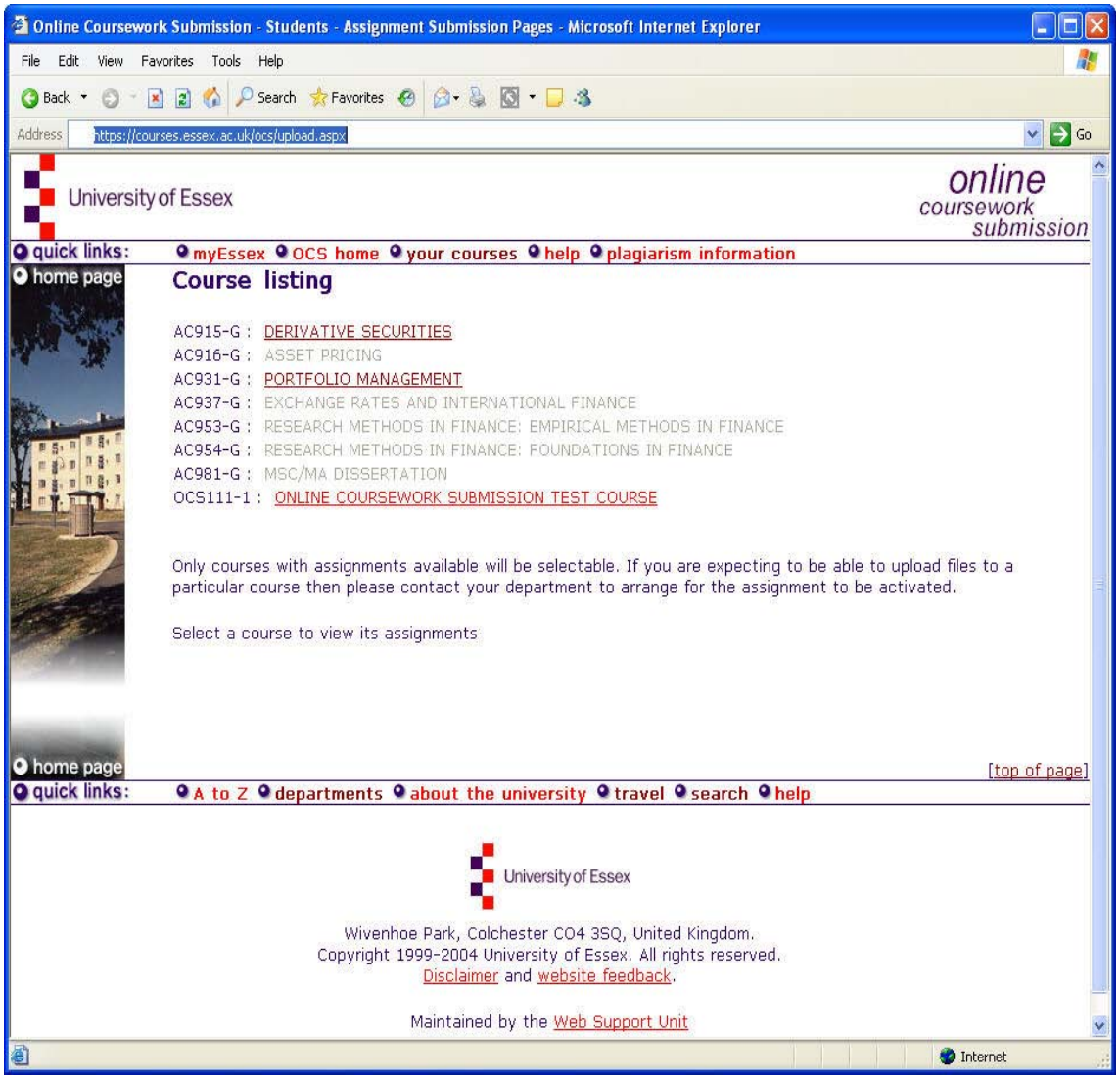
Progress and Challenges

In the pilot phase, 2004-05, the OCS Project developed a range of Web content to support the OCS, this included a dynamic test directory, to enable staff and students to practice uploading files; a set of Help pages; an About section, that explained the functionality of the OCS system; and an interactive plagiarism tutorial for students. The outcomes of the pilot, which ran in eight departments, found that students were generally comfortable with the idea of remote submission, expressing very strong support for it. Staff feedback was also, on the whole, positive, although a number of issues were apparent, most notably on departmental processes and communication between administrative and academic staff, on issues associated with anonymity and departmental policy on deadlines.

In 2005-06 the development of OCS into a university-wide system kept all of the central features that existed in the pilot but was enhanced in terms of user interface, with a style consistent with the Essex University corporate layout, and, in technical terms to meet those elements involved in MIS integration and watermarking as described in the second section of this paper. There also had to be a significant increase in the complexity and presentation of the reporting available to administrative staff.

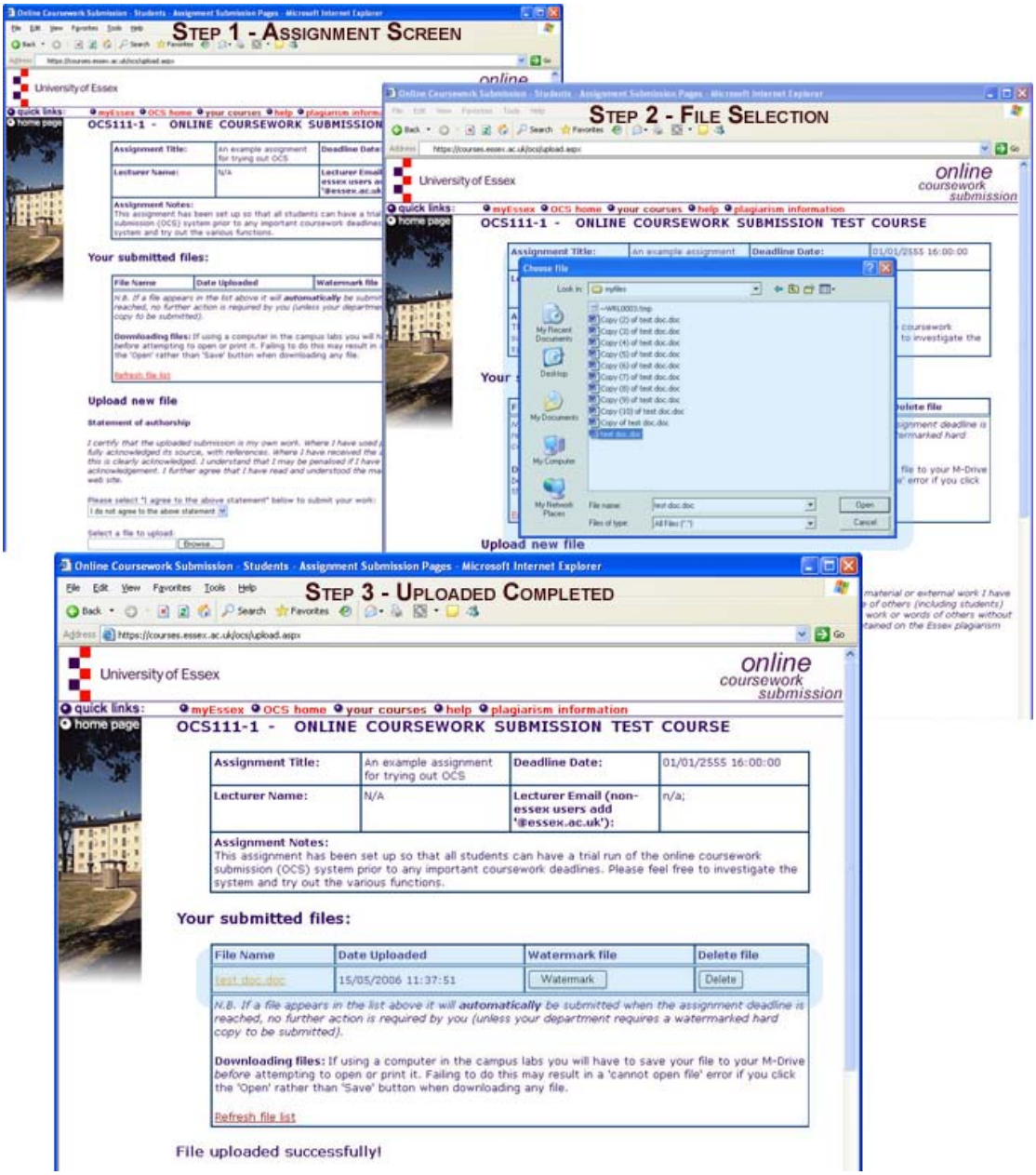
Student Interface

The student interface was overhauled for final release. Whilst there were no serious objections raised to the pilot interface, it was a departmental rather than an institutional design and as such the decision was taken to reformat the presentation to match the university's corporate pages.



Most students will enter the OCS system via the university's student web portal, myEssex. The portal site checks the OCS database to obtain a list of assignments for that student's course list and provides dynamic links directly to the assignments page for that course. Students can also visit the OCS web front end directly should the portal site be temporarily unavailable.

To upload a file for a particular assignment is a simple three step process.



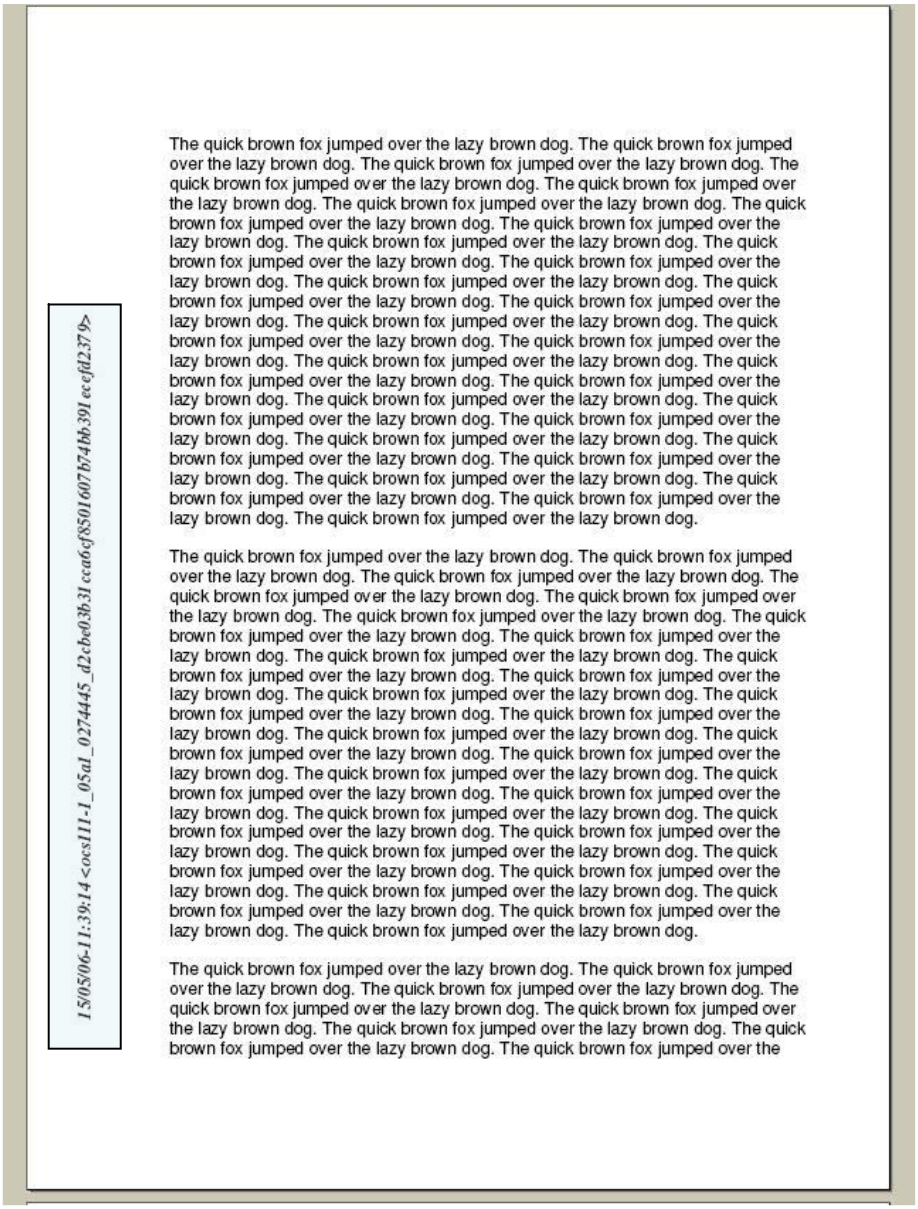
From myEssex they select the course link for which they wish to submit. Secondly they confirm a statement of personal authorship and select the file they wish to upload. Finally they click on the ‘Upload file’ button and complete the process. Watermarking, where necessary, requires a single mouse-click to send the request which is then processed as described below.

Watermarking

Several options were explored when determining how best to achieve this, including some external software solutions, but these were deemed either unreliable or financially non-viable. In the end a combination of two separate windows services, running on the data-store server, handle the watermarking process.

When a user requests a watermark copy of a document they have submitted to OCS, it is renamed according to a specific convention and copied to a directory watched by the *OCS printer service*. When this service detects a new document it opens it (the service uses .Net's interoperability with MS Word to handle the process and allow for greater customisation of output) and prints it to a PDF file using a third-party PDF printer driver. This printer driver automatically outputs to a preset folder, which is watched by the second service in the process – *Watermark watcher*.

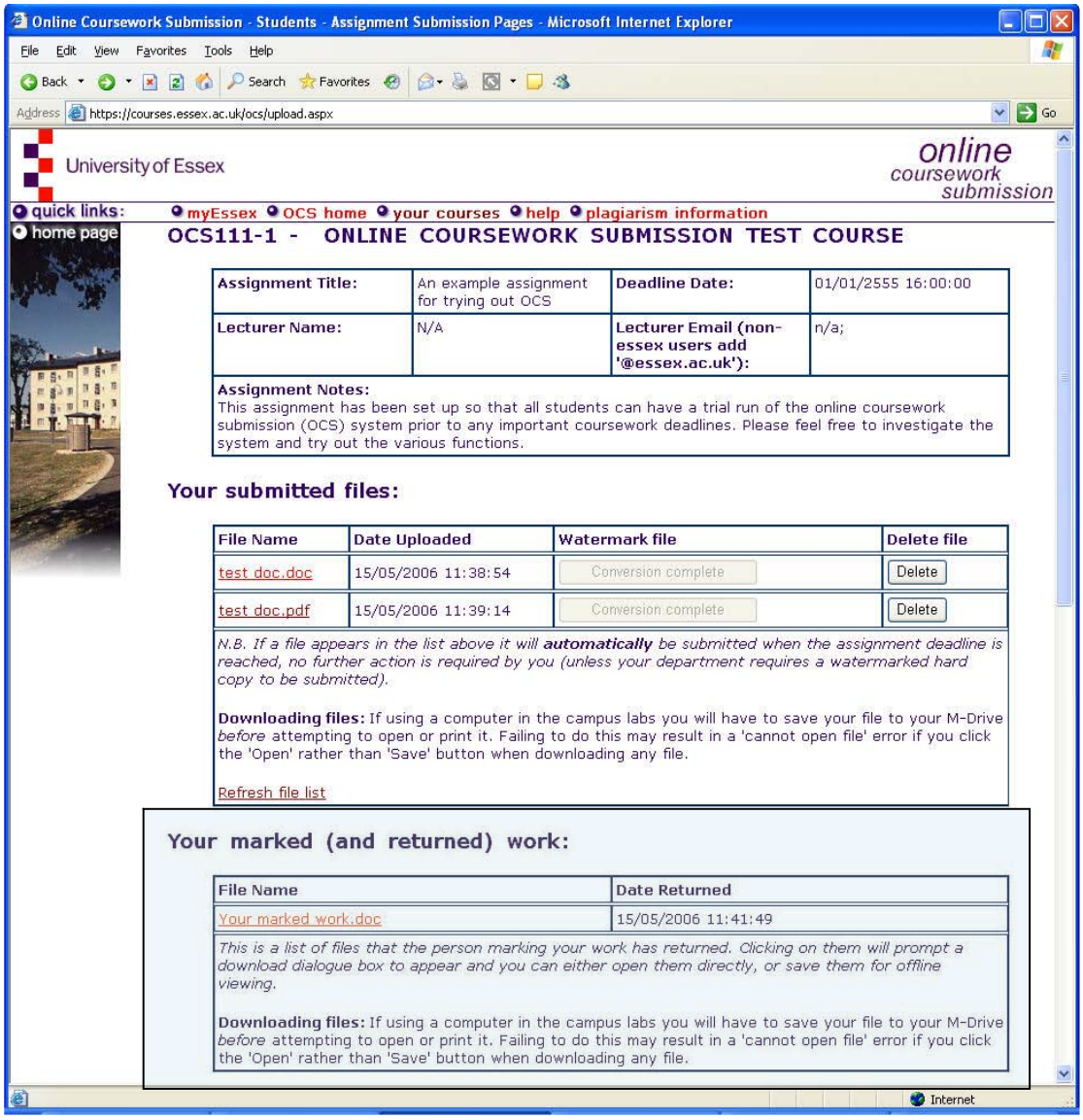
When *Watermark watcher* detects the presence of a new file in the output directory it parses the filename and queries the OCS database to determine which course/assignment/student directory the file should be returned to. On a successful move of the now watermarked file back to the users' directory a confirmation email is sent to the student to notify them of the completion of the process. This last step is important as the whole process is effectively a giant printer queue serving the whole student population and as such students know that watermarking is not instantaneous and to allow sufficient time for it to complete. In reality the process is extremely quick, most documents take under a second to be produced and process completion is normally somewhere in the order of ten to fifteen seconds



One of the main benefits of this method is that students can complete the whole process from any PC with internet access; they are not forced to use university equipment at any stage and so can continue to work as they would have done prior to the introduction of OCS.

Downloading Feedback and Marked Work

Where departments offer return of marked work online the student is able to access work from the moment it is uploaded to the OCS data store. When returned work is present for a particular assignment an additional list appears below the student's submitted file list whereby students can download, save and print staff comments and marks.



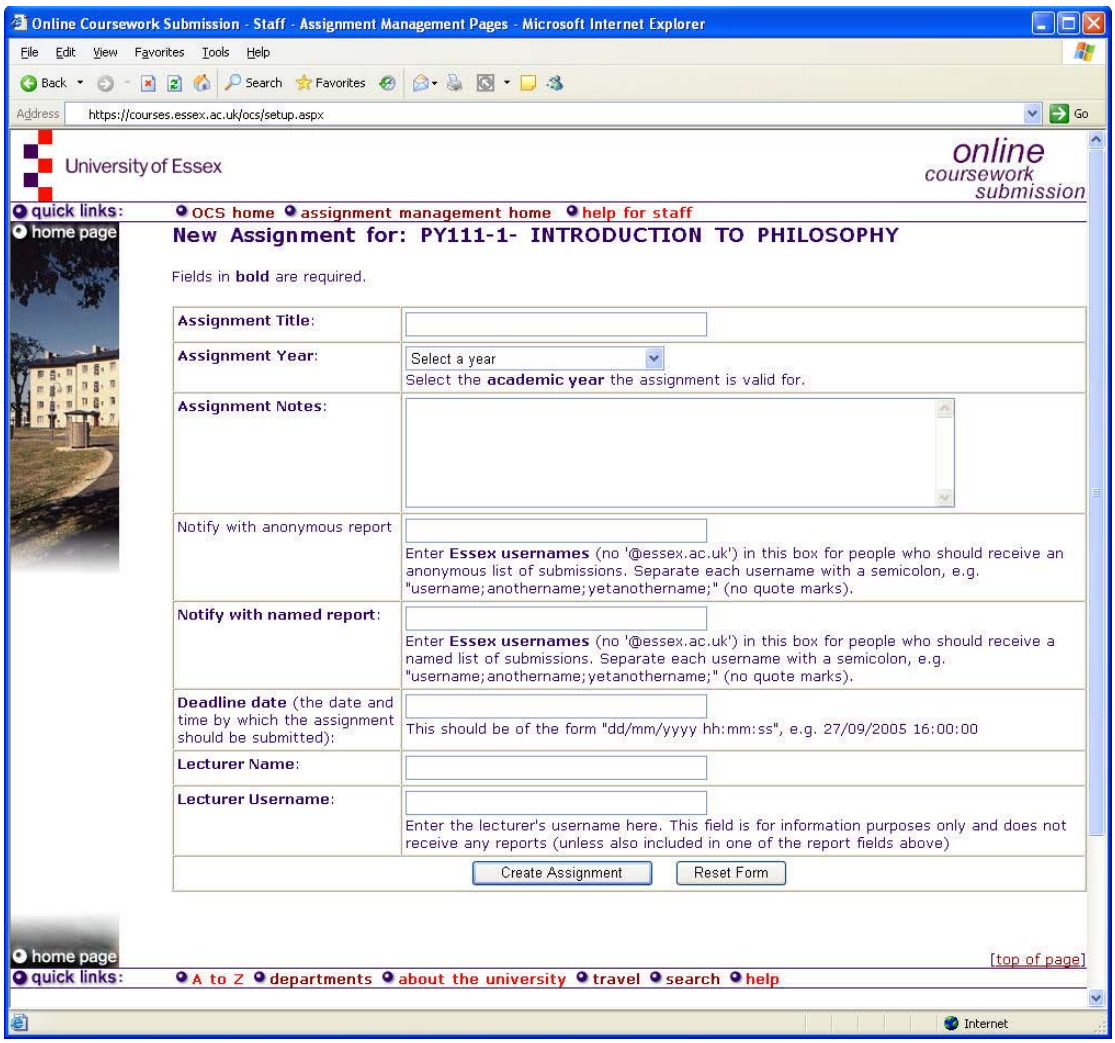
Staff interface

The original pilot contained no staff web interface as setup was administered by one person and assignment zip files were made available via a file share. The final release contains a full interface for staff that allows tiered (as described in ‘anonymous submission’ below) access to the assignment setup, reporting and zip file resources produced by OCS.

Setting up an assignment

As individual departments are responsible for setting up assignments for their courses it was important to devise an interface that was a simple as possible. Setting up an assignment consists in simply going to the OCS management page (the system automatically picks up department and staff status based upon login information and only presents the user with courses relevant to their department), selecting the course for which electronic submission is required and filling in a web form with a brief title, additional notes (which can be as verbose as the department likes) and a deadline date and time. Once this has been done a single mouse click enters the assignment into the OCS

database and from that point onwards any student registered for that course can submit their work.

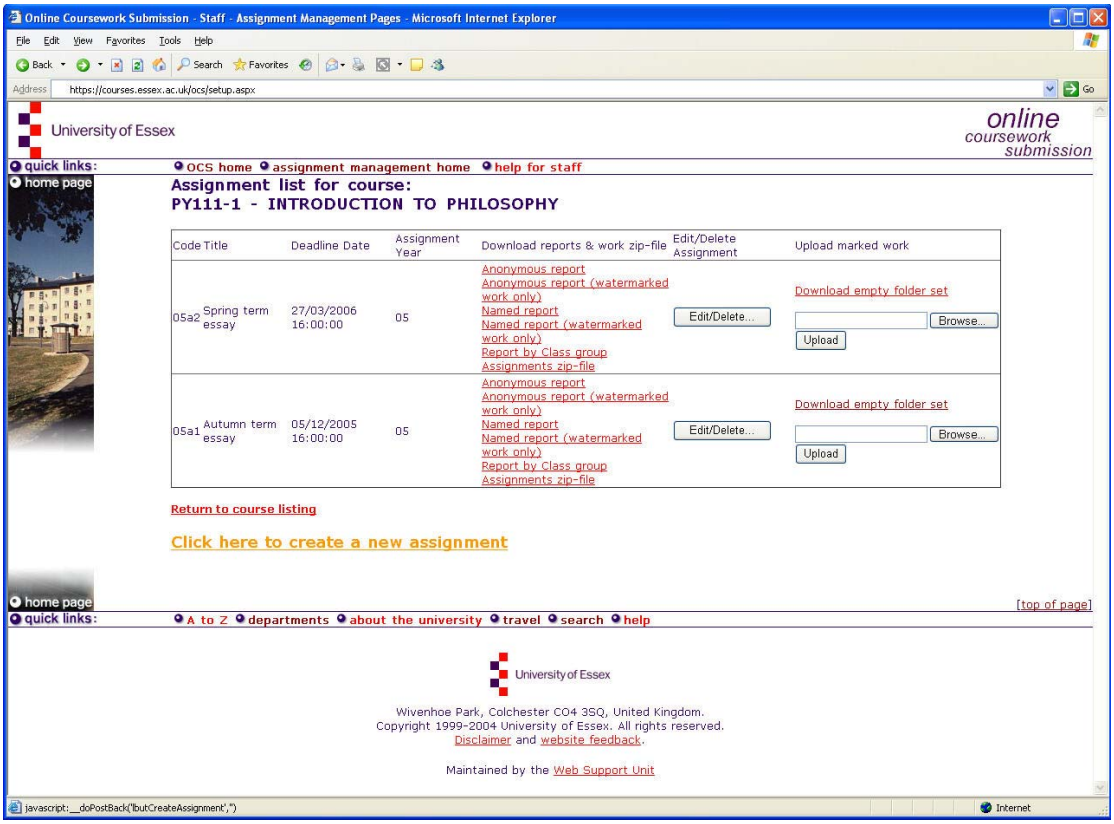


This is a considerable reduction on the administrative burden of the pilot system and requires no technical know-how whatsoever, other than the use of a web browser of course.

If alterations need to be made to a particular assignment staff can easily do so via the same form.

Accessing reports and zip files

This is via the web interface, meaning that staff can access student work and reports for any assignment submitted via OCS from any computer with internet access and a web browser installed. The user simply logs in as normal, selects the course they require downloads from and then picks files from the list contained within each assignment listed for that course.

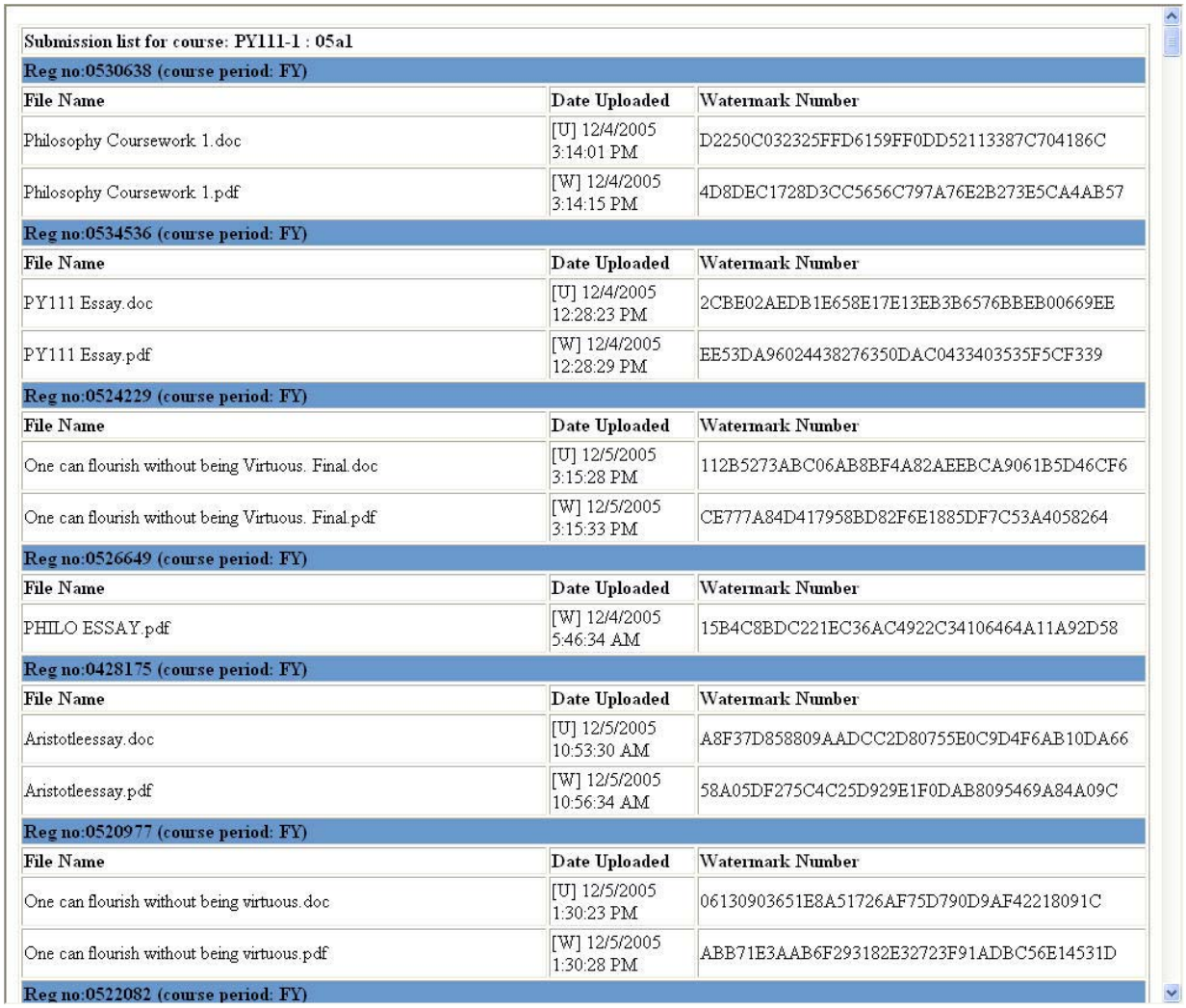


Returning marked work online

The staff interface also includes the facility to return work with comments and marks to students online. Staff marking electronically simply re-zip the student directories extracted from the original OCS zip archive and upload the marked work zip file to the server, which processes it and creates ‘marked work’ directories for those students contained within it. This means that students who have not submitted work for a particular assignment do not see the ‘Returned work’ section and also that staff can return marked work incrementally, returning work as soon as it has been marked where this appropriate.

Reporting

The *Zip sweep service* is responsible not only for producing the assignment zip files that staff download if they wish to mark electronically or submit selected work to the turnitinUK system, but also the production of html reports that can be saved, viewed or printed through the staff OCS management web front end. Reports come in a variety of flavours, anonymous or named, full submission list or just watermarked files and combinations thereof.



Submission list for course: PY111-1 : 05a1		
Reg no:0530638 (course period: FY)		
File Name	**Date Uploaded**	**Watermark Number**
Philosophy Coursework 1.doc	[U] 12/4/2005 3:14:01 PM	D2250C032325FFD6159FF0DD52113387C704186C
Philosophy Coursework 1.pdf	[W] 12/4/2005 3:14:15 PM	4D8DEC1728D3CC5656C797A76E2B273E5CA4AB57
Reg no:0534536 (course period: FY)		
File Name	**Date Uploaded**	**Watermark Number**
PY111 Essay.doc	[U] 12/4/2005 12:28:23 PM	2CBE02AEDB1E658E17E13EB3B6576BBEB00669EE
PY111 Essay.pdf	[W] 12/4/2005 12:28:29 PM	EE53DA96024438276350DAC0433403535F5CF339
Reg no:0524229 (course period: FY)		
File Name	**Date Uploaded**	**Watermark Number**
One can flourish without being Virtuous. Final.doc	[U] 12/5/2005 3:15:28 PM	112B5273ABC06AB8BF4A82AEEBCA9061B5D46CF6
One can flourish without being Virtuous. Final.pdf	[W] 12/5/2005 3:15:33 PM	CE777A84D417958BD82F6E1885DF7C53A4058264
Reg no:0526649 (course period: FY)		
File Name	**Date Uploaded**	**Watermark Number**
PHILO ESSAY.pdf	[W] 12/4/2005 5:46:34 AM	15B4C8BDC221EC36AC4922C34106464A11A92D58
Reg no:0428175 (course period: FY)		
File Name	**Date Uploaded**	**Watermark Number**
Aristotleessay.doc	[U] 12/5/2005 10:53:30 AM	A8F37D858809AADCC2D80755E0C9D4F6AB10DA66
Aristotleessay.pdf	[W] 12/5/2005 10:56:34 AM	58A05DF275C4C25D929E1F0DAB8095469A84A09C
Reg no:0520977 (course period: FY)		
File Name	**Date Uploaded**	**Watermark Number**
One can flourish without being virtuous.doc	[U] 12/5/2005 1:30:23 PM	06130903651E8A51726AF75D790D9AF42218091C
One can flourish without being virtuous.pdf	[W] 12/5/2005 1:30:28 PM	ABB71E3AAB6F293182E32723F91ADBC56E14531D
Reg no:0522082 (course period: FY)		

[N.B. this is an anonymous report]

Anonymous submission

Anonymous submission is not a university-wide requirement yet, but many departments operate anonymous submission using the student registration number as an anonymous identifier. As such all assignment directories are named by registration number (in the pilot the student computer logon was used, but this was not felt to be anonymous enough as it is a composite of first name, initials and last name) and a tiered access system was set up to control access to reports that contain identifying information. Departments are advised to only assign ‘administrative’ access (which includes the setup, editing and deletion of assignments as well as full report access) to central administrative staff and lower tiered access to academic staff (of which there are two varieties, one that only allows access to anonymous reports and the other that allows full report access – in both cases full access to the zip files is possible).

Restricting the number of staff with the highest tier of access has helped to maintain the ‘hub-spoke’ model of administration that many departments operate, whereby any changes to essay deadlines or requirements are fed to students via the administrative team. As noted earlier, one of the issues

raised by the pilot was that of academic and administrative staff communication – one of the benefits of tiered access is to prevent such miscommunication.

MIS Integration

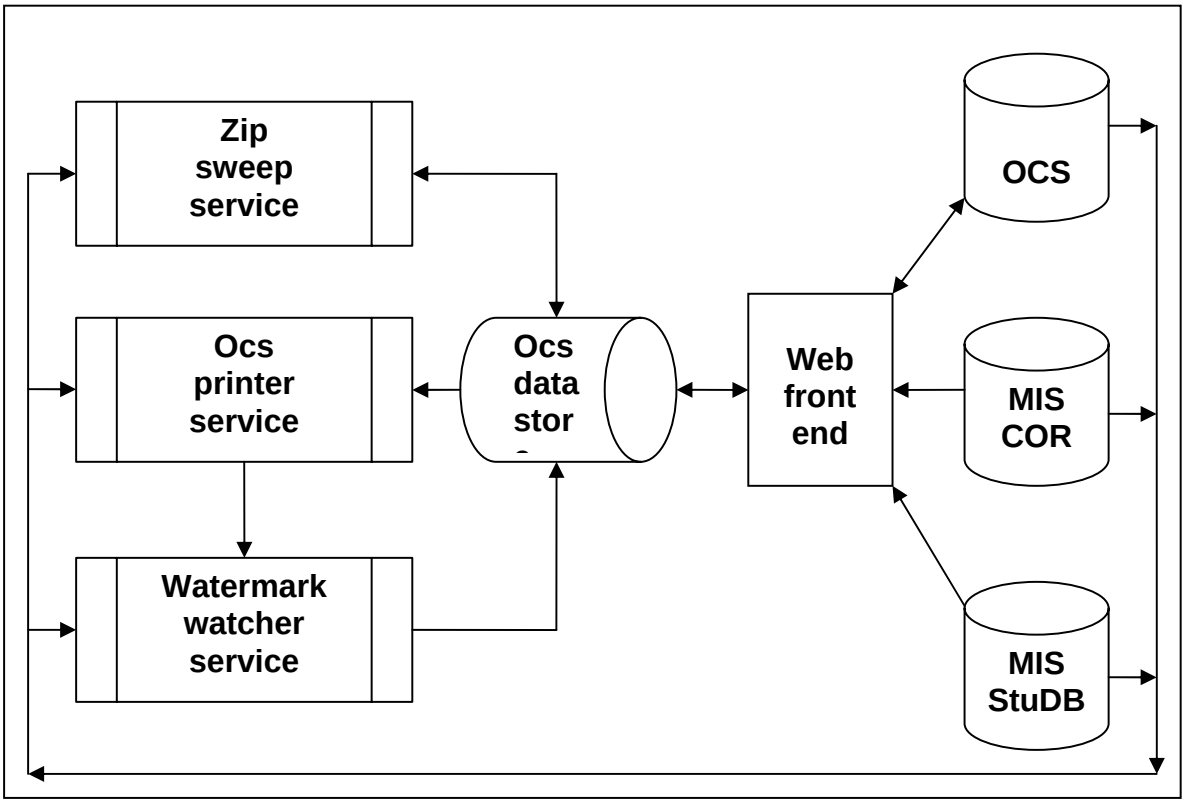
MIS integration was handled by coding the web front end and the backend services to hook directly into the MIS databases that contain the relevant student and class information. Given that the databases are SQL Server and the coding was done in a .Net environment this was especially easy to achieve. This data layer is extractable and can be re-coded to meet future needs and further development as well as export to external institutions, without necessitating a re-design of the presentation layer (the web front end).

Unique Software

One department that expressed an interest in adopting OCS, but that had very specific post-processing requirements, was provided with a special software tool that allowed them to continue to use their own special anonymous number system. The department in question also wanted to remove the burden of sorting through piles of submitted work and decided to take on the print-out of any work for staff who did not want to review and mark electronically.

As such a Windows application was devised that takes the standard zip file output produced by OCS and processes the student work contained within it. The process examines each student folder, opens the submitted work file, inserts a cover and mark sheet automatically and then adds in that student's anonymous number. The whole batch of essays is then sent automatically to a high volume networked printer/photocopier that produces a stack of printed and stapled essays that the administrative staff can simply place directly into the marking staff's post-tray. It also produces a new zip file that contains electronic files that have been made completely anonymous, even down to file author information being removed and replaced by the student's anonymous number.

OCS System Flow Diagram



Key

Zip sweep service: this is a .Net windows service that routinely checks for assignments with deadlines that have past and generates assignment zip files and system reports for staff download via the web front end.

OCS printer service: this .Net windows service operates a PDF printer driver that takes a file requested for watermarking (by a student) and outputs a PDF version of the file with watermarking information inserted within it.

Watermark watcher service: a companion .Net windows service to the OCS printer service; watches the output directory where finished watermark files are sent and re-routes them back to the student and notifies them via email of its availability.

Note: OCS, **MIS COR** and **MIS StuDB** are SQL Server 2005 databases

Conclusion

Recommendations and future plans

The survey on departmental coursework submission in departments, which was undertaken at the outset of the project, ensured that the project was tailored to needs and addressed specific concerns within departments. This approach also helped to engage the buy-in of a number of Departments.

The OCS system has generated some interest from other universities, and the University is looking at ways in which it might develop the OCS more generally for wider uptake across the HE sector. At the University of Essex, the Learning and Teaching Unit will look at the issues associated within online marking, with a view to developing a pilot on online marking via the OCS in the future.

---

[i] VLE Surveys: A longitudinal perspective between March 2001, March 2003 and March 2005 for higher education in the United Kingdom, M Jenkins, T Browne & R walker,
http://www.ucisa.ac.uk/groups/tlig/vle/vle_survey_2005.pdf

EXPORTABLE TECHNOLOGIES: MATHML AND SVG OBJECTS FOR CAA AND WEB CONTENT

Edward Ellis, Martin Greenhow and Justin Hatt

Exportable Technologies: MathML and SVG Objects for CAA and Web Content

Edward Ellis, Martin Greenhow and Justin Hatt
Department of Mathematical Sciences
Brunel University
mapgege@brunel.ac.uk
mastmmg@brunel.ac.uk
Justin.Hatt@brunel.ac.uk

Abstract

The aim of this short paper is to provide an update on our experiences with using Mathematical Mark-up Language (MathML) and Scalable Vector Graphics (SVG) within "Mathletics" – a suite of mathematics and statistics objective question styles written within Perception's QML language/Javascript. We refer here to question **style** to stress that we author according to the pedagogic and algebraic structure of a questions' content; random parameters are chosen at runtime and included within all elements of the question and feedback, including the plain text source for MathML and SVG. This results in each style having thousands, or even millions, of **realisations** seen by the users. Much of what we have developed exists in template files that contain functions called by any question style within the database; such functions are therefore independent of any particular web-based system (we user Perception), indeed, ordinary web pages. We reported on some of these functions at the last CAA Conference (Baruah, Ellis, Gill and Greenhow 2005) whilst basic concepts and terminology for MathML and SVG are introduced by Ellis (2005). It should also be noted that the user's choice of font colours & sizes, and background colour, are all incorporated within the MathML and SVG content. This means that equations and diagrams will be **accessible** to those requiring larger/differently-coloured versions of the content's default options.

This paper further exploits:

The use of tables of arbitrary length to display an algorithm presentation MathML. We here show how MathML can be generated effectively by our "display" functions and incorporated into new question types

SVG diagrams. We show examples of the use of SVG to produce dynamic diagrams and charts that accurately reflect the question's random parameters choice or statistical data. The SVG library of functions produce "objects", such as lines, text boxes, circles, etc that can be called by other functions that build up super-objects such as decision boxes, bar charts, pie charts, Venn

diagrams etc. These are then concatenated within the question, to produce, for example, a flow chart.

SVG graph plotter. Although MathML plotter applets exist, these are generally not open code and therefore cannot be tailored to meet the pedagogic needs of the question and/or feedback. We have therefore developed a graph plotter that gives full control of how any Javascript-defined function is to be plotted, including shading, labelling, highlighting of points of interest such as maxima etc. The utility of such a plotter will be demonstrated within questions.

Content MathML. The test example presented at the last CAA conference has been developed into actual questions.

Another aim of this paper is to include an introduction to our functions. We believe that this will prove useful to a wide range of disciplines that contain mathematical or graphical content. We show how such functions are exploited in an ordinary web page and speculate on the structure of a teacher/lecturer's web page containing printable versions of all our question styles (over 1000) with solutions for each student's realisation. The plan is that the teacher/lecturer will preview a question, select what he/she wants and build up a problem sheet; finally printing will produce, say, 30 realisations of the problem sheet (and matched solutions) for use in traditional teaching settings.

Example 1: The Use of Tables

Figure 1 displays parts of the feedback for a bubble sort question. Note the alert box has been triggered since the input string, although of the correct format, has incorrect length (known from the randomised length of the list of random values, between 1 and 20, given in the question). This is an extension of the checking described in the companion paper at this conference by Baruah et al (2006). The essence of the algorithm is encapsulated in the sequence of feedback tables, where cell colouring is used to show the considered pairs before and after swaps and completed cells (green). The coding for building these tables this is not long and completely general, although for more extensive data sets, the feedback can take too long to render.

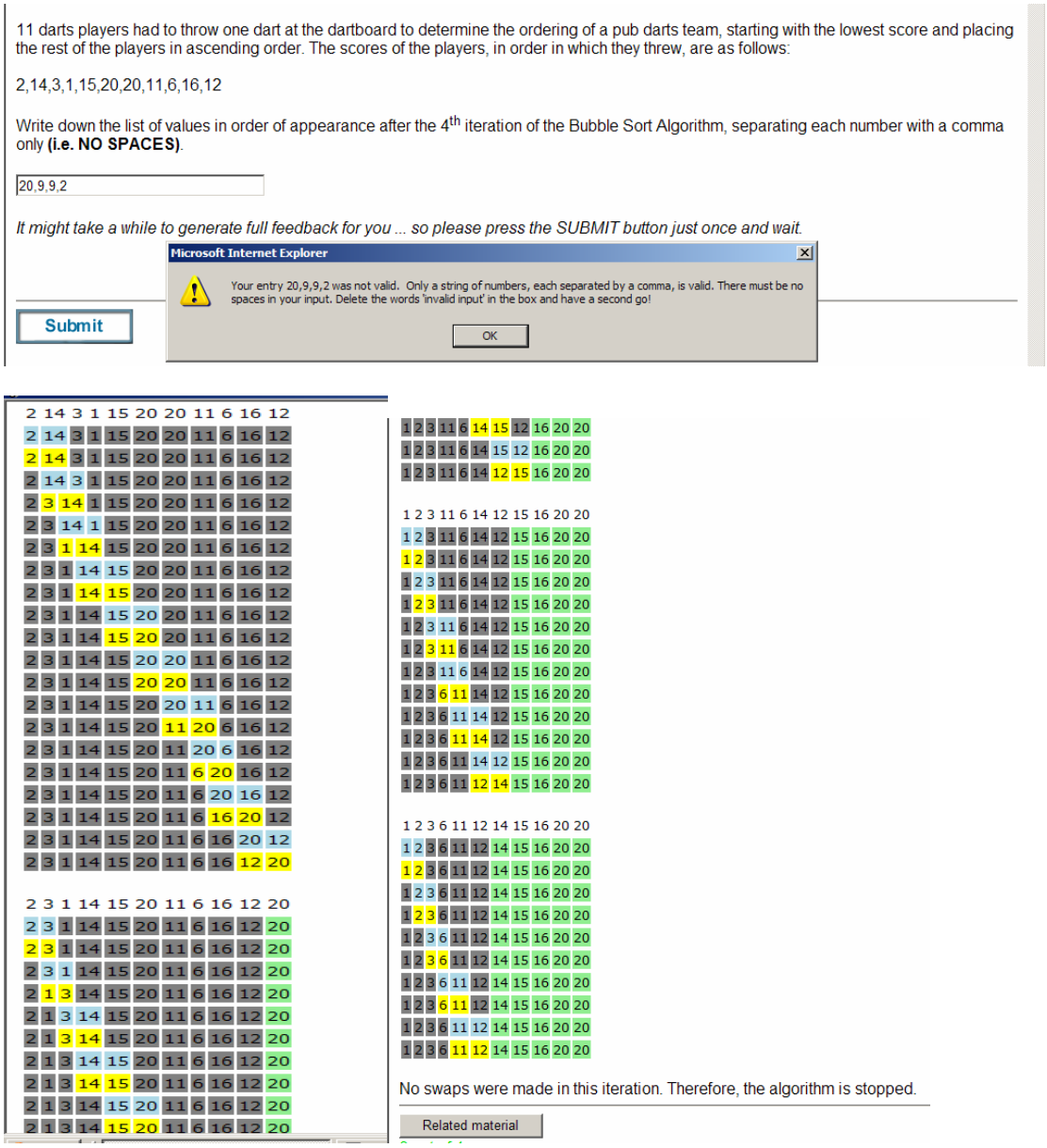


Figure 1. Checking the input validity for a string match question and parts of the feedback tables showing the bubble sort algorithm in action.

Example 2: Javascript and Presentation MathML

By considering a question in linear algebra (LU factorisation) we demonstrate the utility of function to perform calculations and present the results in MathML. The question type is interesting since the required element positions (and question wording) change with each realisation – we call this *positional numerical input* (PNI). Although quite extensive coding is required, the initial set up that guarantees integer values for the answers is quite terse:

```
LT = getrandomtriangularmatrix( random, random, -5, 5, 0,0,1,0 );
//creates the lower triangular matrix.
UT = getrandomtriangularmatrix( random, random, -5, 5, 0,1,0,0 );
//creates the upper triangular matrix.
Bigmatrix = multimatrix(LT,UT);
//multiplies LT and UT together.
```

Here we have essentially started with the answer matrices LT and UT, calling the getrandomtriangularmatrix function:

```
// Function getrandomtriangularmatrix(Nrow,Ncolumn,min,max,allowzero,LU,diagonalones)
will create a matrix of size Nrow x Ncolumn
// elements from min to max and if allowzero !=0 then zero is allowed and if LU=0 then a
Lower matrix is created and
// if LU=1 then an Upper matrix is created
function getrandomtriangularmatrix(Nrow,Ncolumn,min,max,allowzero,LU,diagonalones){
if (LU != 0 && LU != 1){alert("getrandomtriangularmatrix called illegally with LU = "+LU+". This
should be either 0 for a lower triangular matrix, or 1 for an upper triangular matrix")};
var Randomatrix = new Array
Randomatrix[0] = new Array;
for (k = 1 ; k <= Nrow ; k++) {Randomatrix[k] = new Array;}
for (var i = 1; i <= Nrow; i++)
        {for (var j = 1; j <= Ncolumn; j++){
        if (LU == 1){if(i <= j){number = displayarray(1,min,max,allowzero)}else{number = 0};}
          else{
          if (i >= j){number = displayarray(1,min,max,allowzero)}else{number = 0};}
                Randomatrix[i][j] = number;}
        }
if(diagonalones == 1){for(k = 1 ; k <= Nrow ; k++){Randomatrix[k][k] = 1;}}
return (Randomatrix);
}
```

The matrix on the right-hand side (Bigmatrix) is generated by the multimatrix function, i.e. it is correct matrix arithmetic according to this "**reverse engineering**" approach, typical of these questions where one needs to keep control of the complexity of the arithmetic. Certain elements are then overwritten as e.g. $U_{1,3}$ etc before the display matrices are processed by a displaymatrix function that returns the presentation MathML required for rendering. This function (too long to present here) loops round column and

rows to concatenate a returned MathML string that is rendered by the WebEQ viewer applet. Relevant parts of the (shortened) code are:

```
if (a > 0 && b > 0) {
for (k=1 ; k<=rowNumber ; k++) {
        for (i=1; i<=columnNumber; i++) {
                if (k==a && i==b) {rowelements[k] += "<mtd><mi color=RED
        background=YELLOW>" + Rmtrix[k][i] + "</mi></mtd>";}
                else {rowelements[k] += "<mtd><mi>" + Rmtrix[k][i] + "</mi></mtd>";}}}}
for (p=1 ; p<=rowNumber ; p++) {
    therow[p] = "<mtr>" + rowelements[p] + "</mtr>";}
for (f=1 ; f<=rowNumber ; f++) {
    matrixrows[f] = therow[f];}
for (j=1; j<=rowNumber; j++) {
inside += matrixrows[j];}
return inside;
```

We see here the highlighting capabilities of MathML (although figure 2 uses a slightly different technique).

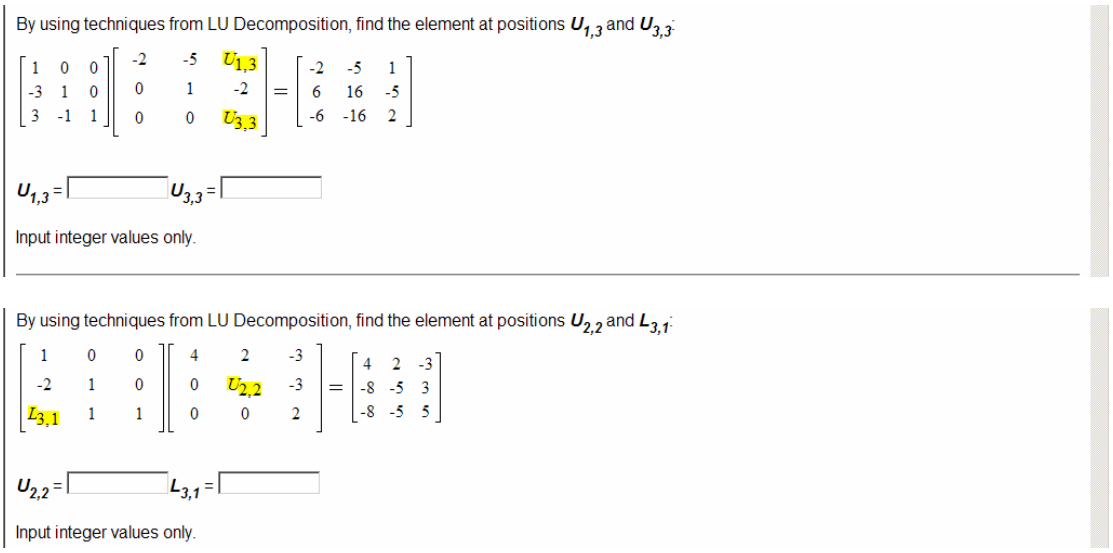


Figure 2. Two realisations of a positional numerical input question.

Examples Using Scalable Vector Graphics

This section looks at the potential of scalable vector graphics (SVG) to enhance the question design or feedback utility. The use of both geometrically accurate diagrams and schematics for mechanics questions has been reported by Gill and Greenhow (2006). Here we look at possibilities in other areas. Figure 3 shows an obvious application, namely geometry.

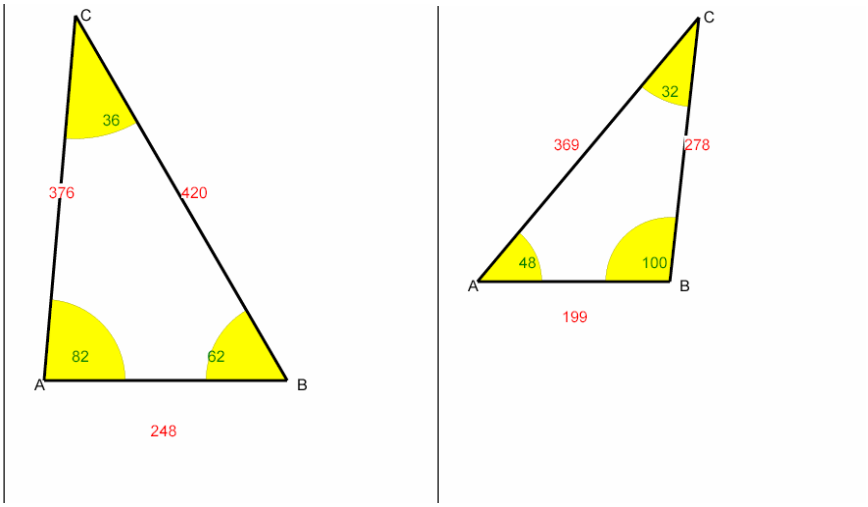


Figure 3. Two realisations of a triangle display.

The coding behind the called function, SVG_triangle, returns the SVG plain text code for rendering by the SVG viewer web page plug-in is quite instructive, but too long to present here. However it is worth noting that the arguments for all coordinates, lengths of sides, angles, labels are all listed, but could be empty strings. This avoids writing many similar functions to handle display where different input data is given. Figure 3 show all arguments, whereas a real question would, for example, omit one of the sides. The SVG_triangle calls functions returning "*atomic components*", such as lines, text boxes, sectors (shown with a yellow background in figure 3) which handle the accessibility features, such as colours and font sizes. Geometric objects such as lines are rescaled according to the font size (both length and line thickness) and use the user's choice of font colour by default. A helper function angle_from_xy(x,y) is also called. It returns the polar angle of point (x,y), needed since Javascripts' arctan function returns the principal value. Finally note that the order of concatenation of the SVG string can be important, see Ellis (2005); in figure 3, the required string order is yellow sectors, then angles, then lines of triangle, then lengths of sides (with opaque background boxes reading the background colour of the page).

Another example of the efficacy of SVG is given in figure 4. The student is asked to apply the first-fit algorithm to the data (the table length, names and weights are randomised). The algorithm produces a shown matrix, but it would be quite natural in class to draw this as a diagram. Dropping the random weights and names into the string-generation loop allows this to happen, producing an accurate and meaningful diagram in the feedback.

In Dropmore County First School's annual tug-of-war competition, 25 contestants have to be placed into groups by the order of their introduction by Reverend J. P. Smythe-Jones-Hamilton, the Vicar. The order of introduction of the contestants, along with their respective weights, are listed in the following chart:

Nadia	Claudia	Julie	Fatima	Lottie	Dennis	Kay	Sandeep	
99	94	115	67	73	71	64	72	
Elizabeth	Alan	Sarah	Asfia	Alan	Clare	Wendy	Stephie	Ingrid
101	88	65	98	71	91	108	84	68
George	Sophie	Lisa	Ajay	Alexandria	Paul	Mary	Daniel	
67	95	74	81	65	86	68	75	

Each contestant's weight is measured in lbs.

The Vicar wants to allocate each group so that its total weight is, at most, 300 lbs (regardless of how many contestants each group has). Calculate the spare capacity of group 6.

[ ] lbs.

Group 1	99	94	67	
Group 2	115	73	71	
Group 3	64	72	101	
Group 4	88	65	98	
Group 5	71	91	108	
Group 6	84	68	67	74
Group 7	95	81	65	
Group 8	86	68	75	

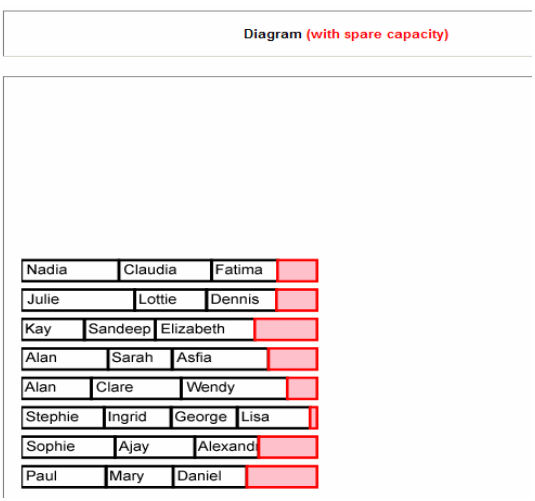


Figure 4. Question stem and components of the feedback for a first-fit question. The SVG diagram accurately displays the data in an effective way.

SVG Graph Plotter

Other developments include an SVG graph plotter. Plotting graphs in SVG has a number of advantages over using either images, or Java Applets. Advantages compared to images have already been covered. The advantages compared to Java Applets is that one can literally draw over the top of the graph. This can prove invaluable in some case. For example, highlighting the roots, turning points, or other significant features of functions.

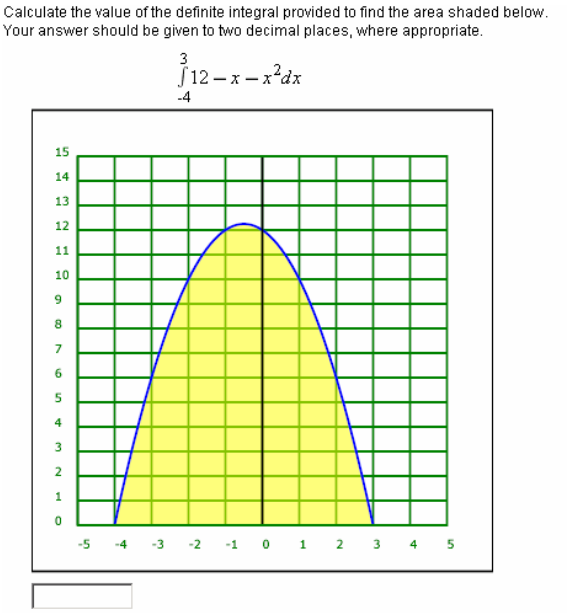


Figure 5. Question stem of an integration question. The SVG graph plots the function to be integrated, according to the random parameters in the integrand and integration limits.

MathML Input (Content MathML)

Content MathML is exploited using a MathML Input question type previously described (Baruah, Ellis, Gill and Greenhow 2005). Entry of free-form Mathematical expressions allows question authors to move away from Multiple Choice questions styles. A great deal of useful information can be obtained from students in this fashion. For example, a question on partial fractions is able to determine the number of fractions the student entered, and the contents of each numerator and/or denominator. Such information can be used to provide targeted feedback.

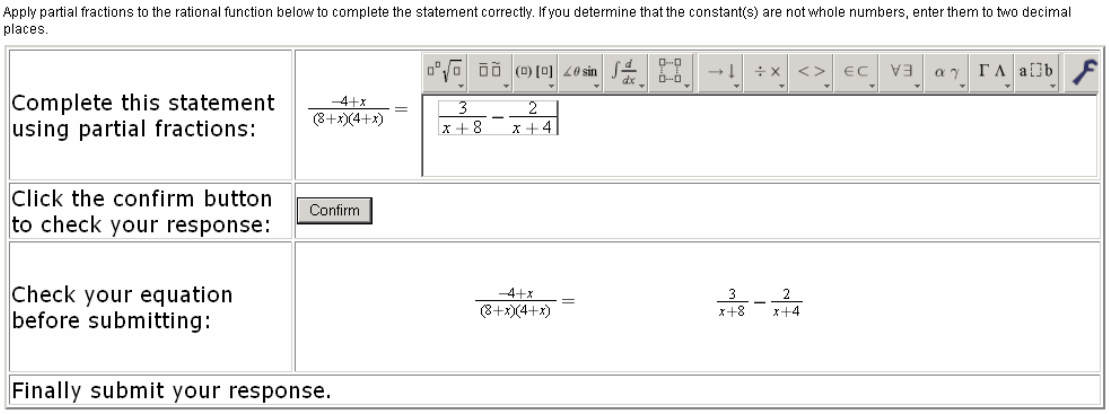


Figure 6. Question stem of a partial fractions question. Use of Content MathML allows detailed analysis of a student's response, without the disadvantages of multiple choice questions types.

A Short List of Available Functions

All of the JavaScript functions can be placed within one of four classes.

1) Generate internal representations of mathematical entities.
2) Manipulate existing internal representations.
3) Convert internal representations into useful alternative representations.
4) Support functions, known as glue.

Examples:

All random generators are in class (1). Examples of these include:

a) rndGraphPoly(degree). This function returns an array representing a polynomial. The polynomial has the property that all turning points exist in the square where x exists [-1,1] and y exists [-1,1]. It is often used in collaboration with the SVG graph plotter.

b) displayarray(num_elements,min,max,allowzero) returns a JavaScript array. That array holds ‘num_elements’ numbers, each in the range [min,max], with the option of excluding zero from that range.

Class (2) is mainly occupied by functions that perform calculations. Examples include:

a) addpolynomial(coeffs1,n,coeffs2,m). This function takes two arrays representing polynomials as arguments. It then returns a new array that represents the sum of the first two arrays.
b) custRound(x,places) rounds the number ‘x’ to the number of decimal places given by places.

Every MathML and SVG generating functions fit in class (3). Example are:

a) displaymatrix(Rmtrix) which returns the presentation MathML representation of the two-dimensional array provided as an argument.

b) SVG_triangle() which takes many arguments. It generates an SVG representation of a triangle, details of which are specified by its arguments. Figure 3 was created using this function.

Accessibility functions and other functions fit in class (4). For Example:

a) femalename(i) returns a female first name from an ethnically balance set of names.

b) getFgColor() retrieves the current foreground setting, stored in the cookie.

A more extensive list will be made available via the MSOR Centre website by the summer of 2006.

Web Page Implementation of the Functions

It is important to stress that all of the above can be implemented in any web based system or indeed, ordinary web pages, such as that shown in figure 6. We believe that extracting the questions' contents to such web pages will be useful for teachers/lecturers who are not able or do not want to use a full CAA system. Whilst marking functionality and answer file writing (and hence analysis) is lost, there are practical advantages to paper-based objective exercise sheets, not least that students can show their workings in the blank spaces to the right of the questions and hand them in.

The anatomy of the web page is quite straightforward: functions are included within script tags in the head, whilst the button, accessibility and credits at the top right of the screen and content is included with a series of question functions in the body (no processing function is needed). Thus a teacher can alter font sizes and colours before printing and randomisation features in question content, including MathML and SVG, is retained.

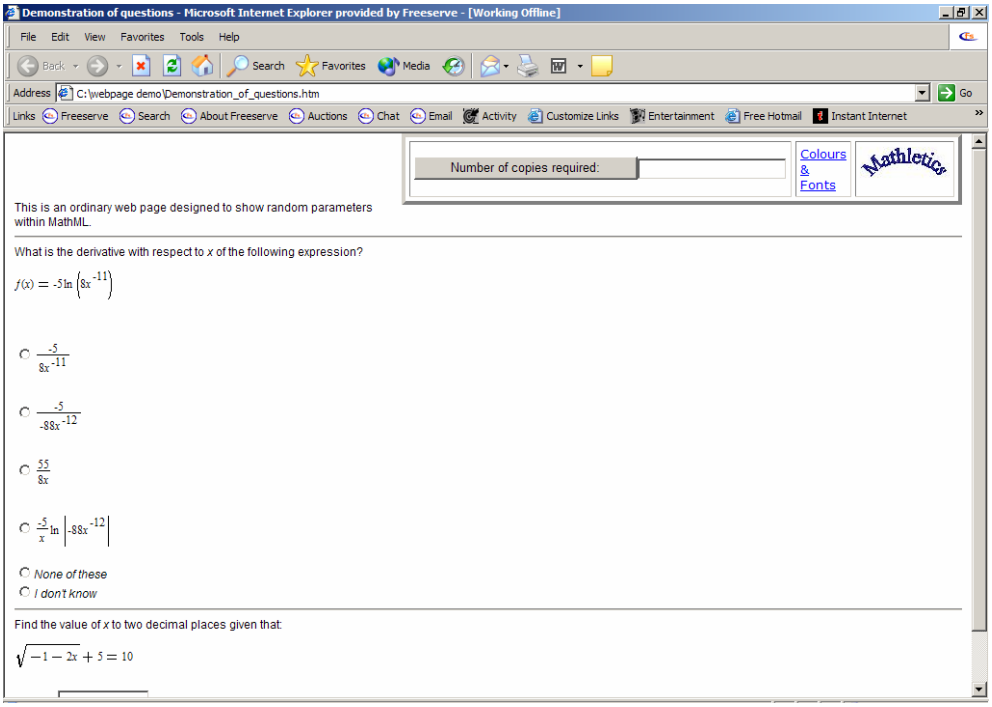


Figure 7. Implementation of functions and MathML in an ordinary web page. The "Number of copies required:" button prints out this problem sheet, reloads it thereby giving questions with new random parameters, prints again etc giving the required number of copies and, separately, numbered answer sheets (planned development).

References

Baruah, N. Gill, M and Greenhow, M 2006 Issues with setting online objective mathematics questions and testing their efficacy *Proc 10th CAA Conf, Loughborough, July*. http://www.caaconference.com/

Ellis, E. 2005 An Introduction to MathML, SVG and JavaScript. MSOR CAA Series. http://mathstore.ac.uk/articles/maths-caa-series/dec2005/

Ellis, E., Baruah, N., Gill, M., Greenhow, M. 2005 Recent developments in setting objective tests in mathematics using QM Perception *Proc 9th CAA Conference, Loughborough, July* http://www.caaconference.com

Gill, M. & Greenhow, M. 2006, Computer-Aided Assessment in Mechanics: what can we do; what can we learn; how far can we go? *Proc IMA Conf Mathematical Education of Engineers, Loughborough, April.*

A COMPARISON OF AN INNOVATIVE WEB-BASED ASSESSMENT TOOL UTILIZING CONFIDENCE MEASUREMENT TO THE TRADITIONAL MULTIPLE CHOICE, SHORT ANSWER AND PROBLEM SOLVING QUESTIONS

Graham Farrell

A Comparison of an Innovative Web-based Assessment Tool Utilizing Confidence Measurement to the Traditional Multiple Choice, Short Answer and Problem Solving Questions

Graham Farrell
Usability and Innovation Group
Swinburne University of Technology
Australia
gfarrell@ict.swin.edu.au

Key Words

Innovative Web-based Assessment Tool
Confidence Measurement
Multiple Choice, Short Answer and Problem Solving Assessments
Comparative Analysis
Convergence Validity
Reliability

Abstract

Computerized assessment is playing a major role in IT education, with extensive utilization of the multiple choice question (MCQ) format. This is mainly due to the ease of adaptation of MCQs into the internet environment, offering extensive advantages to both the student and the instructors. This study analyzes the results of students' grades using an alternative web-based assessment tool and the more traditional modes of assessment, being Multiple Choice Questions, Short answers and Problem Solving (Scenario) questions. The Multiple Choice Questions with Confidence Measurement (MCQCM) is a web based assessment tool that permits the student to register their level of confidence in their answer, and was included as a revision tool for the duration of the semester and as a component of the final exam. Additionally the exam also contained questions using more traditional methods for assessment. A total 43 students sat the final exam producing some interesting results. The statistical analysis indicated that the correlation between the MCQCM and the other alternatives ranges from strong to medium. In addition it appears that the MCQCM demonstrated equal to slightly stronger convergence of validity compared to the traditional MCQ method and the other alternative assessment methods.

Introduction and Literature Review

Educational institutions utilize a variety of assessment options to grade their students and assess the effectiveness and validity of subject content. A critical component of sound educational programs is to assess the learning outcomes throughout the duration of the course, as both a means of giving timely feedback and as a mechanism to grade the students. Black and William (1998) use the term “Assessment” as referring to the group of activities that are undertaken by both teachers and students in self assessment, providing both grades and feedback to modify teaching. Educators appreciate that each kind of assessment has its purpose (Assessment Tools, 2003). It is accepted wisdom that assessment should be an integral part of the learning activities rather than an interruption. (See Principles and Standards for School Mathematics (2000) for example.)

An issue facing educators is what methods of assessment should they be using and what would be the appropriate mix to maximize the feedback and evaluation process? Schuwirth and Van Der Vleuten (2003) state “a well designed assessment program will use different types of questions appropriate for the content being assessed”. The options presently available to the instructors include multiple choice questions (MCQ), short answer questions (SA), longer problem solving questions (PS), case study reports, presentations and other equally effective and proven choices. In the majority of cases the final grade is calculated by combining each separate mark from assessment tasks completed during the subject. The utilization of multiple assessment methods recognizes the need to permit students to demonstrate their knowledge in various methods throughout their learning experience.

Multiple choice questions (MCQs) are highly regarded by instructors (Bacon 2003) and consequently utilized extensively, with world wide experience in their construction (Schuwirth and Van Der Vleuten 2003). In addition, the ease of adaptation to the computer assessment environment has been swift and effective. There are two roles that MCQs play in the balanced educational program. Firstly, MCQs are used extensively as a means of formative assessment (self assessment), where the feedback influences the direction of the students as they journey along their learning path. MCQs are a popular self-assessment option being readily available to the students due to the advancement of technology that now supports its functions. Web based MCQ self-assessment packages permit the student to self assess their knowledge at any time convenient to them, providing instant feedback and in many cases recommended change in directions to their learning path. Secondly, MCQs are also traditionally used for summative assessment for the grading of students, being strategically placed in the exams with various mark allocations directly contributing to the students’ final grade. Their popularity can be attributed to their ability to “yield equivalent reliability and validity in a shorter amount of time” as they have an “economy of scale not found in constructed-response” (Bacon 2003). In addition they are considered to have the ability to test many topic areas in relatively shorter time (Wilson and Case 1993). Bacon (2003) also identifies one advantage of using MCQs is the “Objective” marking as a method of avoiding the “obvious lack of reliability of essay tests”,

as he sites previous work Ashburn's (1938) where subjective marking of short essay answers yielded significant difference in grades when remarked. Schuwirth and Van Der Vleuten (1996) emphasize the growing dissatisfaction with the MCQ format as they rely on recognition of the correct answers, while some see MCQs as only demonstrating knowledge of isolated facts (Wilson and Case 1993). Wilson and Case (1993) also state that they fear this "undue emphasis on recall" will "stimulate students to learn in a like mode". Schuwirth and Van Der Vleuten (2003) go on to recommend variation in the question formats due to the likelihood that students will prepare depending on the types of questions used. Bacon (2003) discusses at length the concerns of some that the MCQ format is too simple and does not assess the complex levels of knowledge, in particular the higher levels of Bloom's (1956) taxonomy of educational objectives (Knowledge, Comprehension, Application, Analysis, Synthesis, Evaluation). Bacon (2003) does recognize the examples of MCQs in Blooms (1956) work that demonstrate the application of MCQ testing designed to assess outcomes at every level. It is also recognized that this level of MCQ is difficult to construct. However, some educators argue strongly that research has demonstrated that the question format is of limited importance and that the construction of the question is critical (Schuwirth and Van Der Vleuten 2003).

The Short Answer (SA) assessment format has equal popularity as the MCQ alternative. Short answer assessment strategies can offer more flexibility, with greater ability to test creativity and higher levels of Bloom's (1956) taxonomy of educational objectives, as outlined previously. However, SAs are resource intensive when grading and are subject to poor reliability due to subjective marking.

The longer Problem Solving (PS) questions are often included in the final exam as it permits the instructor to assess the highest of Blooms levels. The format of these questions usually present the student with a scenario situation which requires the student to call upon many aspects of the subject material to analyze, synthesize and evaluate, offering alternatives in some situations. These are clearly more difficult to grade consistently as there is often not a prescribed correct solution but a number of equally valid alternatives.

In this study we introduce a fourth assessment option. The students are required to complete a formal assessment task utilizing the MCQCM, contributing to their final grade. The MCQCM is a web-based assessment that has been developed over a period of years designed to permit the student to register their confidence in each of their choices and consequently be rewarded or penalized proportionally. (Farrell, Leung, 2004) The MCQCM format is similar to the MCQ display where each question has a stem followed by four options (Klohe 1995, Frary 1993). Once the student commits to an answer ("level") they are required to register their confidence in that choice ("strength"). (Bandara 1983, Betz & Hacket 2002)
Each option of the question must be committed to either correct or incorrect.
The confidence is registered as a %, with 100% stating complete certainty in the choice and a low % representing extreme doubt. Fig 1 demonstrates the tool in action.

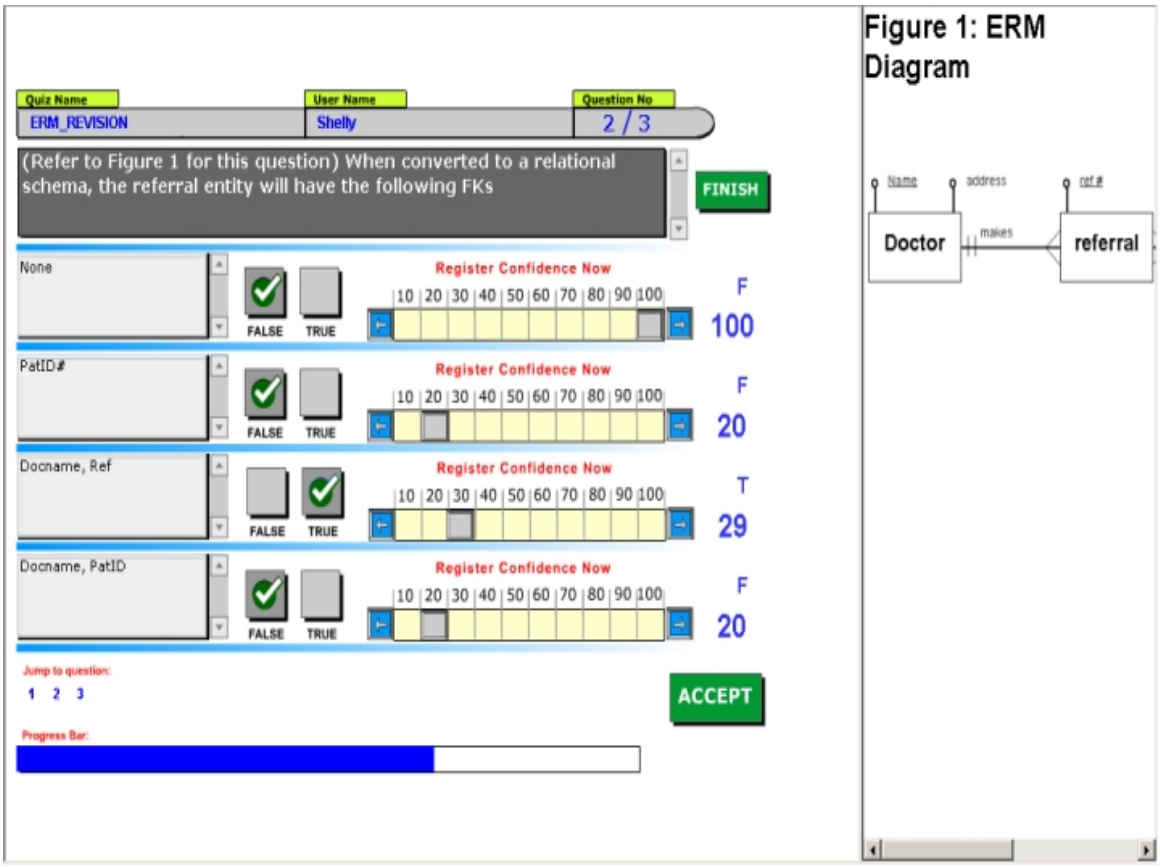


Fig 1: Screen shot demonstrating the tool in use. In this case the ERM is given on the side and the student is required to identify the Foreign Key. This example demonstrates very little confidence by the student in the subject material.

Scoring

Registering a high level of confidence for a correct answer results in a high positive score. (Eg. 100% gives 10 marks), decreasing in increments of 1 for less confidence (90% gives 9, 80% gives 8 etc).
In comparison registering a high % for an incorrect answer gives a large negative result with the same increment (Eg. 100% gives -10, 90% gives -9 etc).
Importantly the students utilize the system as a formative assessment option during the semester and are familiar with the functionality and scoring mechanism.

The Validity of any testing method is mainly assessed using comparison with other test methods (Schuwirth and Van Der Vleuten 1996), yet is often a point of debate (Bacon 2003). Schuwirth and Van Der Vleuten (2003) define the validity as “whether the question actually tests what it is purported to test”. A recognized method of assessing validity is by comparing the correlations between methods of testing that are supposed to measure the same construct (Bacon 2003).

In addition, the Reliability of any testing method is defined as the accuracy of which a score on a test is determined, or more precisely, a score that a student obtains should indicate the score that this student would obtain in any other given (equally difficult) test in the same field ("parallel test") (Schuwirth and Van Der Vleuten 2003).

In previous study (Farrell & Leung 2005) it was demonstrated that the MCQCM provided a rich formative assessment tool, guiding both student and instructor to areas of concern in the student's learning path. The student using MCQCM is not only able to alert the instructor to any areas where knowledge is lacking or incorrect (as in MCQ's), but can also demonstrate areas where they have partial knowledge and/or lack confidence in their knowledge. While the MCQCM proved to be beneficial in its feedback objective it remained to show that it was at least equivalent in its convergence of validity as an assessment tool to the standard accepted MCQ format.
This paper will firstly present an examination which includes four separate methods of assessment. It will then statistically compare the results for each student across each method. A discussion and conclusion will follow to determine the validity of MCQCM as an assessment tool.

Method and Objectives

A total of 43 students sat the final exam as part of the formal grading process of an IT subject.
The exam consisted of an 8 Multiple Choice Question (MCQ) section followed by 8 MCQCMs, 8 Short Answer Question (SA) section and a 2 part Longer Problem Solving questions (PS). The students sat the final 3 Hr exam at the same time on campus. The MCQ and MCQCM sections carried 20% each of the final exam grade, the SA section carried 33% while the longer PS section the remaining 27%. The author of the exam was mindful of Bloom's (1956) taxonomy of educational objectives when constructing the questions to facilitate the assessment of various levels.

The results were collected on the completion of the exam and each question's mark was carefully recorded for analysis.

Results and Discussion

To facilitate this study we investigated the exam results of a cohort of 43 Information Technology students enrolled in the optional subject.

Section	Average Grade	Standard Deviation
MCQ	73%	17.7%
MCQCM	67%	21.0%
SA	85%	9.8%
Problem Solving	75%	14.5%

Table 1:Means and Standard Deviations for each of the sections of the exam

On analysis of the data in Table 1 it is noted that the average grades for all sections of the paper are close, as too are most of the standard deviations. It is observed that the SA section has the greater average grade with a smaller Standard Deviation. Instructors would be quite pleased with these outcomes at this stage.

On further examination and analysis of the data it was found that in most cases there appears to be a good relationship between each of the grades allocated for each of the sections for the individuals. (In a few instances this is not the case) Again this is very pleasing for the instructor as there appears to be a good convergence for each of the assessment areas under consideration. As educators we rely on a reasonable convergence of the grades for each of the sections. Failure to achieve this might indicate poor question construction in a particular section. In this case there does not appear to be any one area of concern.

At this stage, a statistical analysis is appropriate to identify the true relationship between these results.

The correlation for the scores for each of the sections was used to test the convergent validity, using Spearman's Rank Order correlation test.

Due to the number of pairs for comparison the results are displayed in Table 2:

Correlations

			MCQ	PS	MCQCM
Spearman's rho	PS	Correlation Coefficient	.235		
		Sig. (2-tailed)	.129		
		N	43		
	MCQCM	Correlation Coefficient	.436(**)	.302(*)	
		Sig. (2-tailed)	.003	.049	
		N	43	43	
	SA	Correlation Coefficient	.447(**)	.442(**)	.544(**)
		Sig. (2-tailed)	.003	.003	.000
		N	43	43	43

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Table 2 Correlation table for the sections of the exam

The following observations can now be discussed. All of the levels of correlation are as defined by Pallant (2005) reference to (Cohen 1998))

Firstly, let us consider the correlation between the MCQCM and the other sections of the exam paper.

There is a reasonably strong correlation between the MCQCM and the SA section ($r=.544$, $n=43$, $p<.01$).

MCQCM also has a medium correlation with MCQ and PS ($r=.436$, $n=43$, $p<.01$ and $r=.302$, $n=43$, $p<.05$) respectively).

These statistics confirm that there is a convergence of validity for the MCQCM and all of the other sections of the exam. Additionally, these correlations gain strength when considering the Cronbach's Alpha reliability coefficient for the results, demonstrating the internal consistency of .692, (slightly below the recommended minimum of 7.0).

Further, it is interesting to see that the grades for the MCQ section demonstrate a medium correlation to SA ($r=.447$, $n=43$, $p<.01$) and a small correlation to PS ($r=.235$, $n=43$, $p<05$).

SA and PS has a large correlation ($r=.442$, $n=43$, $p<.01$).

Discussions and Conclusions

In conclusion, this study has identified a convergence of validity between MCQCM and all of the other sections of the exam paper, with the strongest correlation being between MCQCM and SA. This observation is very encouraging as the MCQCM was primarily designed as a formative assessment tool to support the learner along the learning path (Farrell& Leung 2002).

Interestingly, the traditional MCQ section of the paper has medium correlation with the SA but only has a small correlation to the PS section. Hence, whilst there is convergence of validity between MCQ and SA there is no significant convergence of validity between the MCQ section and the PS section. This means that a good performance in either section would not predict a good performance in the other.

As a result of these initial observations MCQCM appears to be a valid assessment option, producing grades that have equal reliability as the more traditional methods of assessment. However, MCQCM does not appear to offer any great advantage over the rest of the methods of summative assessment. The question then must be asked, why bother?

Previous investigative work in using MCQCM as a formative assessment tool (Farrell, Leung 2005) has proved that utilizing MCQCM can be highly beneficial to both the student and the instructor as its feedback is often reflective of their confidence in their knowledge of a particular subject material. This often influences the learning path of the individual to address the areas of concern, encouraging management of the learning by the student. (Farrell, Leung, 2005)

This study encourages the utilization of the MCQCM as a summative testing option in the future. It is proposed that the tool continue to be utilized as a formative assessment method for the duration of the semester and be included as part of the final exam, producing more data for analysis. In addition the authors intend on gauging the students' acceptance or rejection of MCQCM as a standard method for summative assessment.

References

Ashburn, Robert (1938). An experiment in essay-type question. *Journal of Experimental Education 7 (1): 1-3*

Assessment tools for Assessment, Evaluation and Curriculum Redesign workshop: *month 7 http://www.thirteen.org/edonline/concept2class/month7/index_sub2.html (Last accessed Aug 2003)*

Bacon, Donald R (2003): Assessing Learning Outcomes: A Comparison of Multiple-Choice and Short-Answer Questions in a Marketing Context: Journal of Marketing Education. Vol 24. No 22. Sage Publications

Black. P and William D (1998): Inside the Black Box: Raising Standards Through Classroom Assessment. Phi Delta Kappan October 1998· Volume 80· Number 2 P 139-149 http://www.pdkintl.org/kappan/kbla9810.htm

Black. P and William D (March 1998): Assessment and Classroom. Learning *Assessment in Education, Vol 5 March* P. 7-74

Bloom, Benjamin S (1956): Taxonomy of educational objectives, hand book 1: Cognitive domain. *New York: Longman Green.*

Bridgeman, Brent, and Charles Lewis (1994): The relationship of essay and multiple-choice scores with grades in collage courses. *Journal of educational measurement 31 (1): 37-50*

Cohen, J. Statistical Power Analysis for the Behavioural Sciences. Hillsdale, NJ: Erlbaum. (1988)

Farrell, G and Lung, Y:Designing an Online Self-Assessment Tool Utilizing Confidence Measurement. Conference Proceedings *IFIP 8.4 WG (2002) P.525-537*

Farrell, G and Lung, Y:Improving the Design an Online Self-Assessment Tool Utilizing Confidence Measurement. Conference Proceedings *Web-based Learning (2002) P.149-159*

Lambert W.T. Schuwirth and C.P.M. van der Vlueten (1996): Quality Control: Assessment and Examinations: *http://www.oeghd.or.at/zeitschrift/1996h1-2/06_art.html (Last accessed Aug 2003)*

Pallant. J. SPSS Survival Manual, 2nd Edition, Allen & Unwin (2005)

Principle and Standards for School Mathematics (2000): *National Council of Teachers of Mathematics - Standards 2000 Project Chpt 2*

Wilson, R. B. and Case, S. M.: Extended Matching Questions: An Alternative to Multiple-choice or Free-response Questions: *Journal of Veterinary Medical Education. Volume 20:3. http://www.utpjournals.com/jour.ihtml?lp=jvme/jvme203/ExtendedMatchingQuestions.html (Last accessed Aug 2003)*

Foreword

Welcome to the tenth International CAA Conference at Loughborough University.

As I look back over the history of the event I feel honoured to have worked with so many enthusiasts, experts, commercial attendees and delegates. I do feel that together, through this event, we have made a significant contribution to the field now widely known as e-Assessment.

CAA Conference continues to attract, develop and host the largest body of continuous research into e-Assessment that I am aware of. This year we have a full two day programme of quality double blind refereed papers from a wide and diverse representation of stakeholders. I would like to thank our Advisory Panel for their contributions in guiding the selection of papers for 2006, and creating what promises to be an excellent and intellectually stimulating event.

Within this book of proceedings you will find contributions from Awarding Bodies, Higher Education, Research Committees, National Projects, along with international initiatives from America, Australia, Canada, Holland, Poland and the Republic of Macedonia.

Topics reported are diverse, ranging from the delivery of mobile assessments, accessibility issues, item banking, service oriented architecture approaches, national case studies, and much more. This year the programme has been loosely ‘themed’ to give an indication of the areas being covered and my thanks go to John Sargeant for assisting with this aspect.

Last year the Joint Information Systems Committee (JISC) announced a number of national initiatives in the e-Assessment area and this year return on Day Two to report progress.

There is genuine cross sector activity in the e-Assessment area, but I feel that divergence prevails. We have a great deal to learn from one another. For example our commercial partners are responsible for some of the largest high stakes e-Assessment activities occurring globally. As an attempt to encourage knowledge transfer I have included abstracts of the commercial presentations in this book and I hope delegates will take the opportunity to benefit from their experiences.

I offer a special thanks to QuestionMark who this year have generously sponsored our Steam Train evening event and conference bags.

I strive to improve our reputation, improve the quality of our content and disseminate our findings. To this end I am very pleased to announce that a collaboration between CAA Conference and the Taylor and Francis journal ‘Assessment and Evaluation in Higher Education’ (AEHE) has been agreed.

The Joanna Bull Prize for best paper will be announced at the event and publicised on the conference web site (www.caaconference.com).

Lastly, a plea to our following of repeat attendees - do please engage with those new to CAA Conference. With your help I'm sure we can generate a stimulating atmosphere of information exchange (and have some fun)!

Enjoy the conference.

Myles Danson, Conference Director
July 2006

A COMPUTER-ASSISTED TEST FOR ACCESSIBLE COMPUTER-ASSISTED ASSESSMENT

Gill Harrison and John Gray

A Computer-Assisted Test for Accessible Computer-Assisted Assessment

Gill Harrison and John Gray
Innovation North
Leeds Metropolitan University
Headingley Campus
Leeds
LS6 3QS
g.harrison@leedsmet.ac.uk
j.gray@leedsmet.ac.uk

Abstract

This paper describes work in progress on the development of a computer-assisted test to be used for staff development purposes. The aim of the test is to raise awareness of disability issues particularly in relation to the use of technology and of CAA, and to include within itself some simulation of the experiences of people with impairments.

Introduction

Making University teaching staff more aware of disability issues is becoming an increasingly important priority in the light of ongoing Government legislation (HMSO, 1995 and 2001). Staff Development material in this area is readily available, for example in the form of Staff Packs from TechDis, but the work presented here relates to a slightly different approach in that the training material is presented in the form of a computer-assisted test. The idea behind this was that staff might have difficulty in finding the time to attend half- or full-day workshops, whereas they might find a short test of half an hour or so to be both manageable and perhaps also enjoyable. The test was intended to give some simulation of the experience of disabled students, as well as to impart information. It was decided that such a test could be developed with fairly modest resources, and that these could be made available within the "Centre of Excellence for Teaching and Learning – Active Learning in Computing" (CETL ALiC), a HEFCE-funded collaborative project involving Leeds Metropolitan University and the universities of Durham, Leeds and Newcastle (HEFCE, 2005, Durham University, 2006). The CETL's objectives include staff development and disability issues.

Aims of the Test

The test questions were developed within Innovation North (the faculty of information and technology) at Leeds Metropolitan University with the initial intention of promoting staff awareness within the Faculty of disability issues, with particular reference to the use of technology in support of disabled students and of the need to consider accessibility in the context of computer-assisted assessment (see Ball, 2005 and Phipps and McCarthy, 2001). Given that the staff and the students they teach are within the computing and technology disciplines, there is a particular emphasis on the role of technology in teaching and learning, as the students need to engage with a range of hardware and software in conjunction with the subject matter of their course. Computer-assisted assessment is widely used, and appears to be popular with both staff and students, but staff appreciation of the need to consider accessibility in this area is not necessarily high. The aims of the test were broadened over time to consider the staff development needs of less technologically oriented staff in other faculties.

Test Content

A strategy for developing questions was devised. Principles of good practice in question design (for example as propounded by Bull and McKenna, 2004) were followed, and advice on content was received from the Disability Services Manager at Leeds Metropolitan University. Because of the aim to simulate certain experiences of disabled students in order to encourage participants to acquire empathy with them, some of the test questions had to be fitted with “escape routes” to allow the questions to be re-displayed or re-spoken without the imposed constraint, or they might be impossible to decipher. This led to a decision to have information available on the screen about how to reformat the question, together with information about the type of impairment illustrated and suggestions about good practice to be followed.

A grouping of four areas was used:

- visual impairment
- hearing impairment
- physical / motor impairment
- cognitive / learning impairment

Ideas for questions within these areas were generated. Sometimes the content of the question and the form of the question could be related, sometimes the content might be unrelated to the form, being either relatively trivial or about an unrelated aspect of disability issues. Some questions required escape routes, whilst others did not.

An analysis of the question ideas was thought to be helpful, to allow the developers to review the overall balance of the test, and a grid was drawn up showing for each question:

- which of the four areas it fell within
- what specific impairment was addressed by the question's form
- what specific impairment was addressed by the question's content
- whether an alternative format was needed (an escape route)
- whether the question related to disability specifically in relation to CAA, or to disability in relation to technology generally

The production of questions proceeded as follows. For each question, an idea was suggested, a storyboard design produced, the question implemented as a Web page using HTML and JavaScript with Cascading Style Sheets, and an analysis according to the factors listed above carried out.

Two Example Questions

1. A question to illustrate colour-blindness has stem "What is the most commonly occurring form of colour-blindness?" and options "Red/Green", "Yellow/Blue" and "Purple/Pink". These options are initially rendered almost illegibly with little colour discrimination between lettering and background, to try to give an experience of how they might appear to a colour-blind person. It is then possible to request that the options be rendered in contrasting colours, with the "Red/Green" option appearing as red text on a green background, and the other two options appropriately coloured. A further request can be made to have the options shown without colour, as black text on a white background. Advice reminding staff to be aware of colour-blindness when using CAA is also displayed, together with relevant references (for example Waggoner, 2004).

2. To draw attention to the difficulties some users experience over fine motor control of a mouse, one question has the radio button option choices moving around the screen whenever an attempt is made to click on them. Again, advice and appropriate references (for example WebAIM, 2006) are offered to staff.

Usage of Test

The test was designed to be used as part of the regular Staff Development sessions for staff teaching in the computing and technology areas, who are the most likely people within the University to be making use of computer-assisted assessment. The current version of the test is planned for first delivery in June 2006 to a group of these staff, and additionally to a group of staff in Leeds Metropolitan's Carnegie Faculty of Sport and Education.

Conclusion and Future Work

There are several issues arising from this work that will need to be studied in the future. They may be summarised as:

- evaluation and improvement of the test itself
- consideration of its context and usage
- range of applicability

Feedback will be sought from those taking the test, and an evaluation of its performance will be carried out. The possibility of receiving input from disabled students, using focus groups, has been discussed with the Disability Services Manager. Improvements to questions, the removal of unsuitable questions and the creation of further questions will be an ongoing process.

Issues regarding the context of such a test need to be thought through – for example, if some of the takers of the test themselves have impairments (possibly undeclared) what would be the effect on them?

The test could be made available at events that are organised from time to time by the CETL to disseminate results from projects undertaken. These events take the form of informal displays (”roadshows”) to which passers-by drop in for information or advice, or of organised workshops. Perhaps the test might be useful to staff teaching in settings other than Leeds Metropolitan University – for example in Further Education Colleges or schools?

In conclusion, this small study has sought to investigate the possibility of using computer-assisted testing as an aid to staff development in the area of disability issues with respect to technology and CAA. Further work, including an evaluation of its success, remains to be carried out.

References

Ball, S. (2005) *A Checklist for Inclusive Assessment: Getting Started with Accessibility* http://www.caaconference.com/pastConferences/2005/proceedings/BallS.pdf (22nd February 2006)

Bull, J. and McKenna, C. (2004) *Blueprint for Computer-Assisted Assessment*. London and New York, RoutledgeFalmer

Durham University (2006) *Active Learning in Computing* http://www.dur.ac.uk/alic/ (May 5th 2006)

Higher Education Funding Council for England (2005) *Centres for Excellence in Teaching and* Learning http://www.hefce.ac.uk/learning/TInits/cetl/ (22nd February 2006)

HMSO (1995) *Disability Discrimination Act 1995 (c. 50)* http://www.opsi.gov.uk/acts/acts1995/1995050.htm (13th April 2006)

HMSO (2001) *Special Educational Needs and Disability Act 2001* http://www.opsi.gov.uk/acts/acts2001/20010010.htm (13th April 2006)

Phipps, L and McCarthy, D. (2001) *Computer Assisted Assessment and Disabilities* http://www.caaconference.co.uk/pastConferences/2001/proceedings/i1.pdf *(5th May 2006)*

Techdis *Staff Packs* http://www.techdis.ac.uk/resources/sites/staffpacks/index.xml (5th May 2006)

Waggoner, T.L. (2004) *Colorblind HomePage* http://colorvisiontesting.com/ (10th May 2006)

WebAIM (2006) *Motor Disabilities* http://www.webaim.org/techniques/motor/ (10th May 2006)

DEVELOPMENT OF A STUDENT-SEARCHABLE DATABASE OF VETERINARY MCQS WITH EDUCATIONAL FEEDBACK, FOR INDEPENDENT LEARNING

S Head and C Ogden

Development of a Searchable Database of Veterinary MCQs with Educational Feedback, for Independent Learning

S Head
Royal Veterinary College
Hawkshead Lane
Hatfield
Hertfordshire
AL9 7TA
shead@rvc.ac.uk

C Ogden
Speedwell Computing Services
The Old Granary
Irthlingborough Road
Wellingborough
NN8 1RG
Cathy@speedwell.co.uk

Abstract

The OCTAVE Project aims to provide the students of the English veterinary schools with a database of Multiple Choice Questions (MCQs) to support their learning. The questions have been written by veterinary academic staff and practitioners, and contain educational feedback to aid the students' understanding of the correct response. There is always a problem when assembling a database to serve multiple institutions in that the curriculum content and sequence is likely to be different. Therefore, it is essential that the students can select the appropriate categories of questions to use.

In order to make the database readily searchable, questions have been meta-tagged so that students from any institution can make selections in a defined subject area. The search tags for the questions are:

- Stage of course: i.e. preclinical or clinical
- Species
- Body system
- Discipline- scientific/clinical
- Sub-disciplines

The search occurs as each tag is chosen and the number of questions available after each search is indicated. This allows students to decide whether they want to focus

the search further or whether they are happy to be presented with all the questions in a certain area.

Students can choose to attempt the selected questions in three different ways:

- Assessment mode – correct/incorrect score given only
- Assessment /Revision mode with correct/incorrect indication and running total given, and with individual question feedback available after taking all questions
- Revision mode/instant feedback which is given after attempting each question

Student activity is recorded and students may retake a previous test or may choose to review/retake only those questions which they previously answered incorrectly.

The feedback that is offered to the student has three components:

If the chosen answer was incorrect the feedback:

- explains why that option is not correct
- gives a hint to the correct answer- but does NOT give the answer
- provides a reference for further study

If the correct answer is chosen, the feedback:

- confirms and reinforces that the answer IS correct.
- gives some further useful information (like icing on the cake)
- provides a reference for further study

This form of feedback follows best educational practice in identifying deficiencies of logic, stimulating student reflection, and offering extra references and information as a “carrot” for completion.

The database may also be used by lecturers at each institution in similar modes, or to select questions for use in institutional assessments. The responses given by students for each question are recorded so that subsequent analysis can determine: the effectiveness of the question, frequency of choice of each distracter and the level of difficulty of the question. This will allow lecturers to choose questions of known difficulty to present to students for formative examinations or use in summative examinations.

3072 questions have been authored onto a Microsoft Word Template, peer reviewed, assembled into an Excel spreadsheet, tagged and imported into the Speedwell Database 'WebQuest'. This will be made available to be accessed through the web by authenticated veterinary students at each of the English veterinary schools after testing is completed.

Introduction

In this short preliminary paper, we will detail the different criteria that have been considered when developing a database of MCQs with formative feedback, that are intended to be used for CAA by staff and students in the English Veterinary Schools. For a number of reasons, the database has not yet been formally used by students, so that analysis of student use and the effectiveness of the database as an e-learning tool has not yet been under taken. The reporting capacity of the database and surveys of student use and experiences will be presented at a later date.

As a consequence of the differences in curricula and teaching strengths of the Veterinary schools, and the individual Institutional requirements, it was agreed that the database would be a standalone facility aiming to improve the learning experience of veterinary students and not institution orientated.
Authoring of questions was based around the learning objectives at each institution which were perceived as being different in detail but broadly similar. As well as staff at the institutions, a number of questions were authored by students at one of the partner institutions. Each question was based on a specific learning objective and for the student authored questions these were reviewed by academic staff prior to their acceptance to the database. In addition, questions were commissioned from Veterinary Surgeons not associated with the teaching institutions. However, even in this instance, questions related to specified learning objectives.
At the start of the Project, the only universal acceptable form of MCQ used at the partner institutions was the 'single best answer' format. However even in this there was not complete agreement. One institution preferred one correct out of four options, the other three institutions requested one correct out of five options.
The database currently comprises 3072 questions of the single best answer from five options format. A little more than 300 of the questions contain images. These questions were collected into an Excel spreadsheet, metatagged and imported to the Multiquest (Speedwell Computing Services, Wellingborough) database.
The questions of the database are 'core' for the Bachelor of Veterinary Medicine program (BVetMed). These will be used in a number of ways by students for formative self testing or the database will be used staff of the veterinary schools, as a depository of peer reviewed, validated questions for setting electronically delivered assessments in various formats.
A specific requirement of the project was that the database should be compatible with Questionmark Perception, (QTI XML format), used by one of the Institutions to hold high stakes examinations. Also it had to be compatible with the Institutional emerging Virtual Learning Environments (VLEs) such as Blackboard. Finally, it was required that the questions could be readily converted for use in PowerPoint, so that time controlled formative or summative assessments could be readily performed; this being used by yet another of the veterinary teaching institutions.

Topic-specific Question Search Facility

A major difficulty when developing a common database for multi-institution use is that the user should be readily able to access the questions that they require without referring to a specific course component. In order to make the search facility generic, the questions have been meta-tagged to allow students from any institution to make selections in defined subject areas. After consideration of highly defined tagging

systems e.g. SNOMED CT®, we have been forced to be pragmatic and have adopted a simple but effective system. Each question is metatagged for

- Stage of course: i.e. preclinical or clinical
- Animal Species
- Body system
- Discipline- scientific/clinical
- Sub-disciplines

The content of each list refers only to items that are present in the database, eliminating the possibility for searching for items that are not available. As further questions are added to the database, it may become necessary to add more items into these search lists and the software is designed to readily allow this.
We have designed the search facility to be user friendly, the major choice items - 'Animal Species', 'Body System' 'Discipline', 'subdisciplines' - being viewed within the same screen.
The choice of an item is made by clicking on it's tickbox in the list, although, in some instances, more than 1 item within the list may be chosen. Re-clicking on a chosen item removes it from the chosen items. We have decided to adopt this scheme after discussion with a number of students who are already users of e-media. Dropdown lists or free text are not considered effective for this activity.
After choosing a tag, the computer searches for appropriately tagged questions occurs and the number of questions available after each search is indicated. This allows students to decide whether they want to focus the search further or whether they are happy to be presented with all the questions available in the chosen area(s). Indeed not all topics need be chosen, if only a particular species is chosen, then the presented questions will be of any relating to that chosen species.
Should the situation of 'zero questions available' occur after an item choice, the user will be able to readily identify which choice led to the prompt. The user may alter that chosen item, or any other chosen area so that selected questions may be found.

Staff Use
The database was designed to fulfil a number of requirements of staff from the different institutions, although the overall principle is as a repository of questions that have been previously used and validated for discrimination and difficulty.
Staff will be able to use the database to select questions for use in a number of situations:

- in a high stakes summative assessment, which may be run in Questionmark Perception (Questionmark Computing Limited) or in a VLE such as Blackboard (Inc) or in Multiquest (Speedwell Computing Services, Wellingborough) or in Microsoft PowerPoint in a time-controlled manner.

- in a formal formative assessment, being presented in the formats noted above.

- as an end of lecture brief test, which may, in addition, use an electronic classroom communication system to gain immediate student 'feedback'. The presentation for this use is likely to be Multiquest or PowerPoint, both of which may run without any further interaction.

- for setting a formative assessment for later use by students within a set period for student self-assessment / e-learning.

Student Use

The database of questions is designed through formative feedback to aid students in self -assessment so that they can develop learning strategies which enhance their factual knowledge. In addition, about one third of the questions in the database are not merely memory recall, rather they provide students with information or material in the stem, and the responses require that students analyse, interpret, or make choices about that material.

Students may use the database to search for relevant questions as defined already. Questions are presented with a stem and 5 possible responses, only 1 of which is correct. Subsequent viewing of a particular question will present these 5 possible responses in a different order so that the content of the response is important not its position in the choice list.

Students may use the database to self assess/ e-learn in a number of modes:

- Self-Assessment mode, where the feedback is the correct/incorrect score only.
  In this mode the student attempts the chosen questions in groups of 10, or as many available if less than 10. On completing the questions the score for the questions answered correctly is given. In addition, there is indication of whether each question was answered correctly or not. There is no other feedback, although the questions may be attempted again in retake mode.

- Self-Assessment /Revision mode.
  In this mode the student may attempt the chosen questions but after completion of all questions, the score is given, together with an indication of whether each question was answered correctly/incorrectly. However, in this mode, each question may be reviewed, together with the feedback about the chosen answer, be it correct or incorrect.

- Revision mode/Instant feedback.
  In this mode, immediately after attempting each question there is feedback on the answer chosen. The question may be re-attempted in order to achieve the correct answer before moving on to the next question. A score is not given to the student. However, the first choice answer to that question is recorded within the reporting system of the database for subsequent analysis related to question difficulty.

- Review of previous tests.
  Every use of the database by a student is recorded within the system. This will allow students to re-view and re-attempt questions that they have previously attempted. On choosing a particular previously attempted set of questions, they are presented with a list of question numbers in the order previously attempted and an indication of whether they answered each question correctly or not. They can choose to view and answer the complete set of questions again, or to attempt only those that they answered incorrectly at their first attempt. They can also choose to answer them in one of the modes discussed

above; Self-assessment, Self-assessment and Re-vision, or Revision. The distracters will be presented in a different order from that used previously.

In all modes chosen, student data is recorded for production of reports on student activity, and details relating to first responses made are stored to allow analysis of the question characteristics.

The establishment of a database of focused questions will allow more informal formative assessment in veterinary courses and for the first time self assessment on material chosen by the student. A similar but less extensive and non- searchable web based database of questions exists for medical students at Birmingham University School of Medicine (MedWeb). Cook (2001) has reported that students' final examination marks were closely related to the number (and frequency) of computer marked assessments that students had tackled and the development of this database is intended to give veterinary students this opportunity for self-directed improvement.

Educationally Instructive 'Feedback'

The term feedback has different meanings to different authors and different forms of feedback have different outcomes (Yorke 2001, Gibbs & Simpson 2004). Members of the different Veterinary schools expressed different wishes in relation to feedback. Some considered that the mark that a student obtained was necessary, others considered that feedback required only giving of the correct answer, and another group considered that marks and a commentary on the student's response were needed.

There are different outcomes from feedback, dependant on the type of feedback given. Feedback given as marks or grades alone has been shown to have negative effects on the self esteem of students of low ability (Craven et al 1991, Wootton 2002),whilst Butler (1988) has demonstrated that comments alone, which may be termed 'remedial feedback', improved students' subsequent interest in learning and performance.

We have taken account of the principles for providing feedback that will stimulate student's current learning suggested by Nicol & MacFarlane –Dick, (2004).
The definition of feedback that we have adopted is that feedback is 'correction of errors' (Bruner, 1974), and that feedback must be effective in leading to a change of student behaviour (Yorke 2003). In other words, the student is required to make some kind of response to complete the feedback loop (Sadler 1989).

According to the format in which the database is used by a student, the feedback given may offer some or all of the following:

A) Scores achieved (ie total score for the entire assessment).

B) Notification of the correctness or not of a response (ie each question answered correctly or not).

C) Guidance based on a student's response – i.e. educationally instructive information designed to stimulate the student to think again and re-answer the question.

Educationally Instructive Information

The model that we have adopted for educationally instructive feedback is an amalgam from various authors but strongly based on results discussed by Gibbs & Simpson (2004).
Feedback can perform several functions:

- correct errors
- develop understanding through explanations
- generate more learning by suggesting specific study tasks

This has been interpreted in a practical manner so that for most questions in the database, the feedback gives an:
1) Explanation of why the answer chosen is not correct, and, if appropriate, an explanation of what the chosen answer actually is/ or does.
2) Offers a hint to the correct answer- or suggests an alternative way of thinking. However the feedback does NOT give the correct answer.
3) Provide a reference to where more information can be found. The Reasoning for this is that

- The student obviously missed looking/reading this before. i.e. "You did not use this before. Read this NOW!"
- The students' Understanding of the concepts is poor and if a student re-reads a topic whilst a question is still uppermost in their minds they are more likely to learn.
- The explanation given in the database may necessarily be brief, the reference giving a greater coverage of the material.

This is to make the students think! Rather than merely paste facts/answers onto their memory banks. There is evidence that students learn best if guided rather that just supplied with teacher packaged factual information (Sadler, 1989).

Conclusion

We acknowledge that this database of MCQs is not targeted on what might be regarded as higher order skills and higher order e-learning experiences achievable through interactive multimedia experiences and e-portfolios. However, in many of the basic and clinical sciences, there is a core of underpinning factual information which needs to be assimilated, and which students, staff, and, ultimately, the general public, need some reassurance, through assessment, has been assimilated. The use of computer-aided methods for both formative and summative assessment of factual information frees valuable academic staff time for the facilitation of learning in areas of identified difficulty, rather than the process of assessment of lower order skills. It also frees time for the development of valid and reliable assessments of

both practical skills and higher order thinking skills, such as those related to problem-solving synthesis and extrapolation of knowledge to novel fields.

The authors of the database have tried to achieve two important advances in comparison to existing computer-aided assessment materials. The way in which the assessments have been structured follow good pedagogical principles in relation to either recognising a correct response, or pushing an examinee towards recognising and learning the correct response. The feedback has also been arranged to stimulate further learning, even for those who know the correct response to an individual question. In addition, with increased diversity of students, and more demand for individual preferences to be incorporated into computer-aided assessment formats, this interface allows examinees to choose, in a user-friendly fashion, both the content of the test, and also its structure, in relation to whether it is scored without feedback, or whether feedback is provided, and how it is provided. Some students clearly are focused on obtaining maximum marks, whereas other students are much more interested in understanding why an answer is correct, and other answers are regarded as incorrect. Even where students have particular preferences as to how they use multiple choice examinations, these are likely to vary according factors such as the closeness of a large degree examination, and their familiarity with the subject area. Therefore, the system provides variety, meaning that individual students can choose, on separate occasions, any of the different modes of use that best suit their needs and favoured styles of learning at a particular time.

Further Development

A further 200 plus questions are being finalised for incorporation into the database and suitable questions (with feedback) from a current e-learning program for veterinary students (CLIVE) will be incorporated. However authoring quality instructive feedback is difficult and many MCQs were intended merely for assessment. The use of the database will be evaluated and reported. In addition there is discussion regarding a including a choice for the use of questions which have a component of confidence testing.

Acknowledgements

This OCTAVE Project (Optimising Computer-aided and Traditional Assessment in Veterinary Education) had been funded by the Higher Education Funding Council as part of FDTL-4.

References

Medweb. http://medweb.bham.ac.uk/caa/

Bruner, J.S. (1974) Toward a Theory of Instruction, Cambridge, Mass: Harvard University Press.

Butler, R. (1988) Enhancing and undermining intrinsic motivation: the effects of task-involving and ego-involving evaluation on interest and involvement. British Journal of Educational Psychology **58,** 1-14.

Cook, A. (2001) Assessing the use of flexible assessment. Assessment and Evaluation in Higher Education, **26,** no 6, 539-549.

Craven, R. G., Marsh, H. w. & Debus, R.L. (1991) Effects of internally focused feedback on the enhancement of academic self-concept. Journal of Educational Psychology **83** (1), 17-27.

Gibbs, G & Simpson, C. (2004) Conditions under which Assessment Supports Students' Learning. Learning and Teaching in Higher Education, Issue 1, 2004-2005.

Nicol, D. & Macfarlane-Dick, D. (2004) Rethinking formative assessment in HE: a theoretical model and seven principles of good feedback practice. Enhancing student learning through effective formative feedback. The Higher Education Academy Generic Centre. 3-14.

Ricketts, C. & Wilkes, S.J. (2002) Improving student performance through Computer-based Assessment: insights from recent research, Assessment & Evaluation in Higher Education, **27** (5), 475-479.

Ricketts, C. & Wilks, S. J. (2002). Improving Student Performance Through Computer-based Assessment: insights from recent research. Assessment & Evaluation in Higher Education, Sep 2002, Vol. 27 Issue 5, p475, 5p. A comparison between online multiple choice tests and OMR marked multiple choice tests

Sadler, D.R. (1989) Formative assessment and the design of instructional systems. Instructional Science, **18,** 119 144.

Wootton, S. (2002) Encouraging learning or measuring failure? Teaching in Higher Education, **7,** no 3, 353-358.

Yorke, M. (2001) Formative assessment and its relevance to retention, Higher Education research and Development. **20** (2), 115-126

Yorke, M. (2003) Formative assessment in higher education: moves toward theory and the enhancement of pedagogic practice. Higher Education **45** (4), 477-501.

SYMBOLIC ASSESSMENT OF FREE TEXT ANSWERS IN A SECOND-LANGUAGE TUTORING SYSTEM

Matthieu Hermet and Stan Szpakowicz

Symbolic Assessment of Free Text Answers in a Second-Language Tutoring System

Matthieu Hermet *
Stan Szpakowicz *†
* School of Information Technology and Engineering
University of Ottawa
Ottawa
Canada
† Institute of Computer Science
Polish Academy of Sciences
Warsaw
Poland
mhermet,szpak@site.uottawa.ca

Abstract

We present an approach to Computer-Assisted Assessment of free-text material based on symbolic analysis of student input. The theory that underlies this approach arises from previous work on DidaLect, a tutoring system for second-language reading skill enhancement. The theory enables the processing of free-text segments for assessment to operate without pre-encoded reference material. A study based on a corpus of 48 student answers to several types of questions has justified our approach, helped define a methodology and design a prototype.

Preliminaries

In the field of Computer-Assisted Assessment (CAA), automated processing of free-text material received from students is becoming a necessity. The range of such material may run from single sentences to whole essays. Even as seemingly small a problem as student answers to open-ended questions poses a variety of serious Natural Language Processing (NLP) challenges. It calls for different approaches, depending on the didactic purpose of the exercises. This, in turn, affects the nature of the textual material that can be submitted to automated assessment.

In NLP, there is a conceptual opposition between symbolic and statistical processing. While the first relies on methods of qualitative analysis, the second uses the distribution of quantitative text features to draw conclusions. The latter is unquestionably powerful when annotated reference material is available. This is what the field of Machine Learning calls *training data*, while the actual student material is referred to as *test data*. Assessment based on statistical technologies would mean finding the closest possible match

between the training and test material based on features. Feedback associated with the found reference match—the assessment—would then be sent to the user, in the form of a mark or comments. A major drawback of this approach is the need to have annotated reference material. It usually means a considerable amount of time and effort. Statistical methods are also by definition inaccurate, even if accuracy of over 90% is not uncommon in some language processing tasks.

Symbolic processing, on the other hand, usually relies on hand-crafted rules of analysis. It is not necessary to annotate large amounts of reference material, though crafting the right rules also takes time. Rules are triggered by feature values which tend to be acquired automatically. Performance may suffer if feature value acquisition is burdened with error. Still, it is fair to say that the very nature of the didactic process and natural languages (especially the number of exceptions at the lexical and semantic level) make exact rules preferred to nearly exact statistical methods.

Ideally, a hybrid approach—collaboration between symbolic and statistical methods—would be the best for the successful future of NLP. This is by and large a matter of NLP research, external to the concerns of CAA.

In CAA, statistical, or quantitative, processing has been preferred for, as it seems, two main reasons. The first is the existence of vast amounts of (passed) student essays or completed drills. This *is* a rich archive of problems already solved. The second reason has to do with applicability: coupled with dialogue, authoring and moderation modules, such CAA tools are reliable and work predictably well. The level of performance depends mainly on the volume of annotated material. Such systems make good summative assessment tools due to their good capacity to recognize correctness within well-defined domains.

The distinction between summative and formative assessment is not always clear. If we are to treat them as opposed to each other—a means to enhance skills through qualitative evaluation versus a means to judge skills through quantitative evaluation—building ensembles of annotated corpora rich enough to enable fully informative feedback can become a vast problem. That is because it would imply annotating all answer possibilities, including (potentially unlimited) incorrect material.

This is a rough view, and again, in practice existing systems tend to exhibit a mixture of both approaches. We believe, however, that our considerations raise the question of finding or using symbolic methods to cope with free-text analysis. Conversely, if we are to understand the problem as one of economy of annotated reference material, the question is this: is there a point in the relation between answer expressiveness and the nature of exercises, beyond which no pre-encoded answers are needed to properly perform assessment?

This is where the interest of our project lies. It originated in another project, *DidaLect*, with its strong foundation of theory of second language learning.

There is a trade-off in CALL in general between the need to design generic solutions to enhance the visibility on the marketplace (SCORM [1]) and the need to keep the tools very specific in order to guarantee reliability (Chen *et al*. [5])—this extends to CAA. Our own interest is in specificity for the sake of

demonstration: to find a proper didactic niche to implement successful symbolic solutions to prove the soundness of symbolic free-text processing within CAA or, more modestly, to test its feasibility.

The Problem

DidaLect (Balcom *et al*. [4], Desrochers *et al*. [7]) is an adaptive didactic software designed to enhance the reading ability of French-as-a-Second-Language (FSL) students working autonomously. It is firmly rooted in theories coming from the fields of education, cognition and psycholinguistics. Its Virtual Learning Environment is composed of a placement test, a tutorial and resources which support the acquisition of reading skills, for example dictionaries. *DidaLect* is therefore a good example of so-called eLearning Intelligent Tutoring System. First, the Computer Adaptive Placement Test (CAPT) (Laurier [14]) evaluates the learner on her level of French. Next, the learner is directed to a series of texts of varying difficulty, coupled with a set of comprehension-testing multiple-choice questions. The system selects text difficulty as a function of the CAPT results and the test results for the current text.

The theory behind *DidaLect*'s implementation is of crucial importance to the basic design of our free-text answer processing module, which strongly delimits the nature of questions that the student can be asked. We believe that placing such limitations on question types, assuming a solid theoretical foundation, is half of the job of building an unsupervised free-text CAA module. Very briefly, an important aspect of text comprehension is to understand the communication goals expressed by means of language. Such goals are accessible through cognitive operations of sense acquisition as well as through the awareness one has of these operations. All this is embedded in the common cultural background of the author and the reader (Duquette *et al*. [8]).

Assessment

Our system, yet unnamed, is not intended to mark answers, but rather to provide evaluation to the user on the quality of their material, in linguistic terms and on content in relation to the reference. No matter how good a CAA system is, no such system can cope with so-called bad-faith user material, such as answers correctly formulated, but deliberately crafted to fool the machine. Ellipsis, for instance, is a fine rhetorical way to answer a question, but no system can get its accuracy. So, the role of the lecturer is merely to create questions, which only requires knowledge of question categories in the field of text comprehension.

There are a number of implemented open-text CAA systems, often commercial, such as E-rater [3] and Qualrus [10]. E-rater is an Automated Essay Scoring (AES) system, marking and evaluating essays based on a set of pre-scored essays. Human raters mark training-set essays on content and fluency through the evaluation of variables, to be correlated automatically by the system in order to grant a mark. In real-world situation, E-rater is used in combination with human raters to properly assess essays. Qualrus is presented as an "Intelligent Qualitative Analysis Program". It functions as a toolbox for designing assignments as well as assessment tasks. Its

assessment capabilities are a function of both integrated NLP tools and lecturer encoding of what is to be assigned. This makes it an authoring tool rather than a straight CAA module, but it nevertheless can perform tasks of open-ended question marking and evaluation.

Texts

According to literature on the subject, there are two main types of texts: narrative and informative (Chiasson [6]). Informative texts are supposed to exhibit more complex and varying structure, which makes them more difficult to comprehend; on the other hand, they lend themselves more easily to categorization. All texts in the present prototype of *DidaLect* are informative texts, divided between four categories with fairly balanced membership: description, comparison, cause-effect and problem-solution (Richgels *et al*. [11]). The texts are news articles from general or popular-science publications. A text has normally 1-2 pages.

Questions

The categorization of questions works along two dimensions: the cognition processes needed to build understanding, and the form. Cognitively, there are three main categories of questions, addressing three forms of comprehension: literal, interpretative and critical (Chiasson [6]). It is quite difficult (or perhaps not yet feasible) to automate assessment processes for open-ended answers to questions in the two last categories. We can only realistically deal with literal comprehension questions, which have to do mainly with definitions and causal relations in texts.

Categorization by form recognizes Text-Explicit, Text-Implicit and Script-Implicit questions (Pearson *et al*. [9]). The last of these categories requires that the learners perform inference between the text and their own world knowledge; this makes answers in this category difficult to process automatically. The other two categories allow answer construction by recovering (maybe partially) the necessary fragments from one or a few sentences in the text.

If we retain only the first cognition category and the two first form categories, we believe that the resulting questions lead to open-ended answers which can lend themselves to automatic assessment processing.

- Text Explicit questions: dependence on a single sentence

> […] Comme l'avaient calculé les astronomes, l'année tibétaine 1999 débute le 16 février, lors de la nouvelle lune. **Certaines années, pour contourner des conjonctions planétaires de mauvais augure, les Tibétains suppriment des mois du calendrier ou en ajoutent d'autres**. Dans ce cas, la période du Nouvel An, appelée Lhossar, peut tomber un mois avant ou après, par rapport à notre calendrier occidental. […]

Q: Pourquoi les tibétains suppriment ou ajoutent-ils certains mois au calendrier?

- Text Implicit questions: dependence on several sentences, adjacent or (rarely) dispersed in the text.

In the following example, the sentences are not co-referenced. In such cases, we choose to encode question in two ways, one to be displayed and one to be kept by the system in a “closure” form (“replace *quoi* by the answer”).

> […] Abraham, lui, avait compris qu'il fallait sacrifier son fils à son dieu. Quelle bêtise, dirions-nous aujourd’hui! **Vouloir sacrifier son fils à son dieu. Il faut vraiment être primitif. Et pourtant, je me demande si les sociétés modernes, y compris notre société québécoise, ne sont pas un nouvel Abraham qui sacrifie de nouveaux Isaac à quelques divinités. […]**

Q_display: Comment l’auteur juge-t-il l’infanticide sacrificiel?

Q_machine: Il faut vraiment être quoi pour vouloir sacrifier son fils à dieu?

In a more complex case, the sentences are co-referenced, which enables dynamic tracking of the reference sentences making the answer using co-reference resolution techniques.

> […] Quand survient l'impact, on assiste à une réaction en chaîne: **le détecteur de décélération situé à l'avant du véhicule génère instantanément un courant électrique, qui déclenche une amorce, qui elle-même enflamme un mélange allumeur. Ce dernier met finalement le feu à l'agent propulseur responsable du gonflement du coussin**. Toute l'opération se déroule extrêmement rapidement, soit à 300 km/h. […]

Q: Quelle est la réaction en chaîne qui se produit lorsque survient un impact?

We consider that it is possible to address the issue of assessing free-text answers for such types of questions as long as the original text is known to the system.

Processing

It is a two-phase procedure to automate the assessment of free-text answers to the types of questions such as those presented in the preceding section. The first phase checks the content. It consists in comparing the learner answer LRN with the reference answer REF, represented by the text segment from which the question has been built. The second phase checks the syntactic and lexical form. Actually, the two steps are combined in the sense that content assessment works on the results of form analysis. This design seems logical, because lexical selection shapes the content as much as it affects the syntactic form.

Briefly, the procedure proceeds as follows:

1. Create words lists:
    a. words of LRN absent in REF,
    b. words of REF absent in LRN,
    c. words uncommon.
2. Perform dependency parsing of LRN and REF, producing certain dependency relations among lexical items.
3. Use a dictionary of synonym to identify synonymy between words on lists 1a and 1b.

4. Use the dependency relations from step 2, beginning with those containing synonyms found in step 3, to build trees for both sentences. Building is done by breadth-first search, which maximises the probability of discovering new/different lexical material.
5. When the process halts, trees should be completed, as should be records of any diverging lexical material between LRN to REF.
6. Check the syntax of LRN to verify if it conformity to REF, either by
    a. identity: LRN and REF have same structure,
    b. equivalence: sentence LRN is a syntactic equivalent of sentence REF, using certain pre-encoded equivalence rules.

This procedure allows us to capture student errors as follows:

1. agreement: step 2,
2. orthography: step 2,
3. synonyms: step 3,
4. missing content: step 4,
5. syntax in general: step 5.

This procedure does not yet cope with the evaluation of supplementary material. The problem is that of computing the value (in terms of contents compared with REF) of any kind of supplementary material which a student can put in the answer. At present, we can address this issue only partially by comparing the supplementary segment with the rest of the text from which REF comes. This can be explained by our observation that students tend to mix various parts of the text in their answers. Then, we can use co-reference to judge to some degree the coherence of the addition. This further procedure amounts to answering the following question: does the supplementary material interact with the theme of the question somewhere in text? And if it does, at which syntactic level? This is, however, a somewhat uninformed way of solving the problem, without regard to deep semantics. It is a partial solution which has not been tested yet.

Example

> « Selon Yves Grimard et Serge Tremblay, les précipitations acides agissent sur les écosystèmes lacustres depuis 75 ans, **soit depuis l'essor de l'industrialisation et du transport automobile**. *Au cours du XXe siècle, l'acidité des lacs de l'Outaouais s'est multiplié par 10 environ*, ce qui est trop rapide pour qu'un organisme vivant s'y acclimate. »

The following question is Text-Implicit. In order to link the two sentences needed to relate question and answer fragment (*Italics* and **bold**), the question is also encoded under closure form.

> Q_display: Pourquoi l'acidité des lacs de l'Outaouais s'est-elle multipliée par 10 au cours du XXème siècle?

Q_machine: Depuis quoi les precipitations agissant sur les ecosystèmes lacustres ont multiplié par 10 l'acidité des lacs de l'Outaouais*?*

S1: Depuis l'essor de l'industrialisation et du transport automobile[1].

S2: A cause de l'essor de l'industrialisation et du transport automobile.

Creating word lists for S1 will signal the identity of form, as lists 1 and 2 are empty. The list of words in common contains all words of both chunks REF and S1. In such cases, a mere surface comparison of REF and S1 will suffice to assess S1.

Creating words lists for S2 will yield the following result:

- L1: A, cause, de
- L2: depuis
- L3 ; essor, industrialisation, transport, automobile[2]

There is no synonymy relation between *cause* and *depuis*. But checking *cause* in the synonymy dictionary will enable detection of the compositional form of *à cause de*.

A fourth list is created to record words present in Q_display and absent from the set of words contained in both Q_machine and answer segment. This only yields *pourquoi* which is synonymous with *cause*, as shown by Memodata [2]. We have no means of knowing whether *cause* stands for *depuis*, but at this stage we know that it correctly corresponds to the question marker *pourquoi*. As the system cannot go any further in lexical comparison, it moves to the next step, parsing.

S2, (partial) syntactic analysis using XIP [12]

```
NMOD_POSIT1_RIGHT_ADJ(transport,automobile)
NARG_POSIT1_CLOSED_NOUN_INDIR(essor,de,industrialisation)
COORDITEMS_CLOSED_PREP_NOUN(essor,transport)
PREPOBJ_CLOSED(A cause de,essor)
PREPOBJ_CLOSED(de,industrialisation)
PREPOBJ(du,transport)
0>GROUPE{PP{A cause de NP{l' essor}} PP{de NP{l' industrialisation}} et PP{du NP{transport}} AP{automobile} .}
```

REF

```
NMOD_POSIT1_RIGHT_ADJ(transport,automobile)
NARG_POSIT1_CLOSED_NOUN_INDIR(essor,de,industrialisation)
COORDITEMS_CLOSED_PREP_NOUN(essor,transport)
PREPOBJ_CLOSED(Depuis,essor)
PREPOBJ_CLOSED(de,industrialisation)
```

[1] These are the two answers we obtained for the question. These should show the tendency of students to re-use text chunks.

[2] Function words are discarded from L3.

PREPOBJ(du,transport)

1>GROUPE{PP{Depuis NP{l' essor}} PP{de NP{l' industrialisation}} et PP{du NP{transport}} AP{automobile} .}

As we cannot initiate tree-building starting with synonyms (there are none) and as there is no verb phrase to choose as sentence head, the order is to begin with the first relation in the analysis[3]. Here, it is the same for both:

NMOD_POSIT1_RIGHT_ADJ(transport,automobile)

Tree-building performs as follows.

- Retrieve all relations in which a modified term appears (here, *transport*):

COORDITEMS_CLOSED_PREP_NOUN(essor,transport)

- Merge the relations:

COORDITEMS_CLOSED_PREP_NOUN(essor,
NMOD_POSIT1_RIGHT_ADJ(transport,automobile))

This composite relation here is the same for both sentences. This determines the selection of a word on which to iterate merging. The policy is to select the most promising word in terms of semantic importance, or in terms of the probability of discovering supplementary material. To simplify, the resulting complete composite relations are as follows, getting rid of DET relations:

COORDITEMS_CLOSED_PREP_NOUN(PREPOBJ_CLOSED(Depuis,
NARG_POSIT1_CLOSED_NOUN_INDIR(essor,de,industrialisation)),
NMOD_POSIT1_RIGHT_ADJ(transport,automobile))

COORDITEMS_CLOSED_PREP_NOUN(PREPOBJ_CLOSED(A cause de,
NARG_POSIT1_CLOSED_NOUN_INDIR(essor,de,industrialisation)),
NMOD_POSIT1_RIGHT_ADJ(transport,automobile))

Two conclusions can be drawn from this process and analysis. First, *à cause de* as well as *depuis* have both been recognized at parsing time as prepositional phrase heads. Second, the student neither added nor subtracted any textual material with respect to the reference answer. We know, therefore, that no lexeme has undergone any reformulation and that the sentences have identical syntactic structure. As *Á cause de* has also been recognized as a proper answer connector to why-questions, and as it fits the sentence syntax, S2 will be assessed as correct.

Assessment

The example we followed in section 3 shows no errors. We chose it to keep the explanation short while still describing the processing possibilities. The errors, if any, are captured during processing. We examine in turn all types of errors.

[3] The policy for the selection of relations, in case the system has to choose between several, is to favour higher-order categories (SUBJ, OBJ, REL, COORDITEMS…) over lower-order (NMOD, ADJMOD, PREPOBJ…).

Ortography and Agreement

XIP (Aït-Mokthar *et al*. [13], [12]), our parser, outputs the number and gender of the words in addition to what has been shown. A comparison between the lexical files of LRN and REF is all we need to assess the contents with respect to orthography and agreement. This poses the question of number generalization (*les hommes* can be equivalent to *l'homme*), as a student can choose to use singular for plural in an attempt to generalize number. This problem has been left for future work.

Synonymy

The system can only give a partial judgement on the exact pertinence of lexical reformulation. Synonymy is easy to detect with *Memodata basis*, the synonymy dictionary [2], even across parts of speech. Errors are simply recorded as wrong lexical reformulation choices at given syntactic positions, in comparison to REF. We have no means of evaluating such errors in supplementary material. Errors in prepositions are also recorded at this stage, still using *Memodata basis*.

Content

Once the content correspondence between REF and LRN has been established when building trees, the problem is to know whether LRN contains part or all of REF, or even more than REF. Partial correspondence is detected by modifier or complement gaps in LRN with respect to REF, and can only be signalled to the user. An answer is still considered acceptable if it contains only lexical heads. Supplementary material is evaluated through syntax and through the relation which supplementary elements have to other occurrences of heads in the rest of the text.

Syntax

Syntax is assessed through rules of reformulations as well as through heuristics. Rules of reformulations establish correspondence between structures equivalent in meaning but different in form. Those categories of reformulations include mainly nominalization, passive/active and pronominalization. This is achieved by comparing the structure of LRN and REF. Heuristics detect clause reduction in a procedure supported by lists of attribute, state and action verbs; in clause reduction, a phrase containing a verb or modified nouns is reduced to one of its member. The main idea behind this machinery is that reformulation has recursive power: it can occur at the level of the whole sentence or at the phrase level.

Future Work and Conclusion

To keep the list of future tasks short, we prefer future work to strengthen what has already been achieved rather than adding functionality. That is why our main objective is to have an exhaustive set of reformulation rules and heuristics in order to address typical mistakes that FSL students commit, as observed in a set of fifty 20-page journals written by FSL students.

In the present state, we can recognize 46 answers (to 16 questions) out of 48 answers gathered from students during experiments.

Acknowledgement

This work has been partially supported by the Social Sciences and Humanities Research Council of Canada, in the program "Initiative on the New Economy".

References

[1] SCORM/ADL (Advanced Distributed Learning) at http://www.adlnet.org

[2] Alexandria by Memodata at http://www.memodata.com

[3] Y. Attali, J. Burstein (2006). "Automated Essay Scoring with e-rater® V.2.". *Journal of Technology, Learning and Assessment, 4(3)*. Available from http://www.jtla.org

[4] P. Balcom, T. Copeck, S. Szpakowicz (2006). "DidaLect: Conception, Implantation et Evaluation Initiale". Technologies Langagières et Apprentissage des Langues: Actes du Colloque tenu dans le Cadre du Congrès de l'ACFAS, 11-12 mai 2004, sous la direction de L. Duquette et C. St Jacques, Montréal: ACFAS Cahier No. 105.

[5] L. Chen, N. Tokuda (1999). "A New Diagnostic System for J-E Translation ILTS". *Proc Machine Translation Summit*, 608-616.

[6] J. Chiasson (1990). *La Compréhension en Lecture*. Gaëtan Morin Eds, Montréal.

[7] A. Desrochers, L. Duquette, S. Szpakowicz (2004). "Adaptive Courseware for Reading Comprehension in French as a Second Language: The Challenges of Multidisciplinary in CALL". *Proc 11th international CALL Conference*, Antwerpen, 85-91.

[8] L. Duquette, A. Desrochers (2006). "Appuyer la Compréhension en Lecture à l'Aide d'un Logiciel Adaptatif". Technologies Langagières et Apprentissage des Langues: Actes du Colloque tenu dans le Cadre du Congrès de l'ACFAS, 11-12 mai 2004, sous la direction de L. Duquette et C. St Jacques, Montréal: ACFAS Cahier No. 105.

[9] D. Pearson, D. Johnson (1978). *Teaching Reading Comprehension*. New York, Holt, Rinehart and Winston.

[10] Qualrus by Ideaworks at http://ideaworks.com/qualrus.shtml

[11] D. Richgels, L. McGee, R. Lemax, C. Sheard (1987). "Awareness of Four Text Structures: Effects on Recall of Repository Texts". *Reading Research Quarterly XXII*, 177-197.

[12] XIP Parser, Xerox Research Center Europe Technical Document, 2003.

[13] S. Ait-Mokhtar, J.-P. Chanod, C. Roux (2001). "A Multi-Input Dependency Parser". *Proc. Seventh IWPT (International Workshop on Parsing Technologies)*, Beijing.

[14] M. Laurier (1999). "The development of an adaptive test for placement in French". M. Chalboub-Deville (ed.), *Development and research in computer adaptive language testing*. Cambridge: University of Cambridge Examinations Syndicate / Cambridge University Press, 122-135.

IDENTIFYING EASSESSMENT DEVELOPMENT PRIORITIES THROUGH USER EVALUATION

Susie Hill and Martyn Ware

Identifying eAssessment development priorities through user evaluation

Susie Hill
Research & Information Services
Scottish Qualifications Authority
Susie.Hill@sqa.org.uk

Martyn Ware
Business Manager
CAA
Scottish Qualifications Authority
Martyn.Ware@sqa.org.uk

SOLAR Project: Innovating Assessment in Scotland

The Scottish OnLine Assessment Resources (SOLAR)[1] project is being led by the Scottish Qualifications Authority (SQA)[2] with funding from the Scottish Funding Council for Further and Higher Education (SFC)[3] and the European Social Fund (ESF)[4].

The Scottish Funding Council developed an eLearning strategy in 2003 and revised it in 2005. In the context of Scotland's colleges, the strategy identified eAssessment as an important activity. Separately, SQA developed a strategy for computer assisted assessment (CAA) in 2003, and for eAssessment more specifically in 2005. These strategies sought to highlight the benefits to SQA, its centres and candidates from increased use of CAA and eAssessment and set out some of the ways in which SQA planned to seek to increase their use.

The two major objectives of SOLAR are: to develop summative online assessments for units within Higher National (HN) qualifications[5]; and to provide staff development in the writing and use of these assessments. The project will make an important contribution to the wider programme of work currently underway to modernise the HN qualifications portfolio[6].

[1] http://www.solarproject.org.uk
[2] http://www.sqa.org.uk
[3] http://www.sfc.ac.uk
[4] http://www.esf.gov.uk

[5] Higher National Certificates (HNCs) and Higher National Diplomas (HNDs) are intended for candidates at a post-school but below degree level, and are mostly taken in colleges of further education.
[6] In the late 1990s, SQA launched a review of the design criteria for HNCs and HNDs, and, as a result, a new set of design principles was introduced in 2003. A rolling programme of HN

Development of summative eAssessments for HN is ground-breaking work not only for SQA, but also for the sector. It has the potential to have a significant impact on the nature and type of assessment being delivered in Scotland over the next ten years and on the way in which the assessment process is managed in FE colleges. Therefore, another aim of the project is to help inform the creation of a sustainable model for the development, delivery, and maintenance of eAssessments in the future.

To create the assessments, subject specialist lecturers and moderators from the further education sector are drawn together into small development teams of five or six for each curriculum area. They are then trained by SQA in the authoring, peer review, and moderation of eAssessments; this includes training on how to interpret unit specifications to help in the identification of sections which may be suitable for eAssessment of the type being developed under the project. The activity undertaken by members of the writing teams has been mapped against two of SQA's eLearning suite of qualifications (*Diploma in eAssessment* and *Diploma in eLearning Production*) with the aim of ensuring that they are able to gain as much credit as possible through their participation in the project.

After training, the authors and moderators work in subject teams, each responsible for developing and approving assessments that are fit for purpose in their own particular curricular areas. Teams were initially created within the following curricular areas: Computing & IT; Engineering; Care; Communication; Hospitality; Languages; and Administration & IT.

Subject Team Evaluations

The SOLAR project began in late 2004 and two formative evaluations have since taken place – one in spring 2005 and the other in autumn 2005. These evaluations focused on the experiences of the authors and moderators; not only on the effectiveness of the processes in achieving the aims and objectives of the project, but also their potential to provide a long-term process that could underpin the development and delivery of eAssessments to support HN and other qualifications.

Over the period covered by the evaluations there were more than 40 authors/moderators involved within the project. The spring evaluation was primarily based upon the project review meeting and follow-up interviews. The autumn evaluation was based on an online survey and follow-up interviews and had a lower response rate than the first. However, both evaluations were focused around the same themes: planning and preparation; technical support; prior knowledge requirements; subject-specific issues; funding and payment; the moderation procedure; and the outcomes of the project. The evaluations aimed to identify areas where the processes were working well and, conversely, identify areas where there were still issues to be resolved and processes to be improved upon. As such, we were interested in any

---

modernisation, using the new design principles, aims to achieve a modern, coherent HNC/D portfolio. For more information, please see http://www.sqa.org.uk.

changes in authors' perceptions and experiences between the two evaluations.

Communication

Through the evaluations, it became clear that the effectiveness of communication within the different authoring teams depended very much on previous experiences – in certain areas, members worked well together as they were already colleagues or had experience of working with each other on previous projects; in areas where this was not the case, communication was generally poorer with a consequent impact on the effectiveness of the group.

The evaluation report recommended that the SOLAR project strengthened support for discussion within curriculum groups (either online or face-to-face) and that more workshops and group events involving all curricular teams take place to encourage continuous communication throughout the development process. It was generally agreed that more group events would help enable authors to share their experiences. (However, when a follow-up information sharing event was organised late in 2005, few participants registered).

Working Practices

As a result of the evaluations, we appointed co-ordinators within each of the teams to both ensure that communication improved and ensure that each team had an opportunity to decide the best way for them to work. The evaluation found that teams that were working well together – such as Computing & IT – could share their working practices with other teams, to see if there were lessons to be learned.

Subject Suitability

Some teams in the 'softer' subject areas thought that aids such as working models to copy and adapt would help them develop their assessments, as would exemplars of how different subjects can make use of objective testing. Furthermore, although most participants in the evaluations thought that they had enough prior knowledge before the start of the project, it was evident that the ways in which teams worked differed significantly. For example, for the Computing & IT team, issues were around content, while for some other teams (such as those in 'softer' subject areas) issues centred – and time was taken up – more on technology, ease of use, and the suitability of this type of eAssessment within their subject area more generally.

Many of the participants in the project came with preconceptions about the place and benefit of eAssessment and objective testing within their subject area – these centred around the suitability of the subject, the college culture, the nature of the students undertaking the qualification, and the limitations of unit specification requirements. Participants' experiences within the project have given them the opportunity to challenge long-held beliefs – with positive results. For example, with the area of Communication (traditionally a bastion of resistance when discussing the place of eAssessment), the lecturers

involved found real benefit in considering and challenging these issues. They identified areas where eAssessment would have advantages for the subject, students, and lecturers, but were held back by a combination of curriculum design, college culture, and the authoring technology.

Moreover, most of the respondents in the later evaluation felt that their subject area was suited to objective testing. While objective testing clearly lends itself more easily to some subject areas than others, the authors and moderators involved in the project appreciated being able to explore the issues around it. Indeed, by the second evaluation, a number of authors had changed their view on whether such an approach was suitable for their subject area; the project has the potential to change attitudes to objective testing and eAssessment generally. Importantly, perhaps, the evaluations – especially the second – found that, in most subject areas, the majority of participants thought that the SOLAR project does provide the basis for a sustainable model and framework for future development of eAssessments.

The initial list of areas within which we created teams was based on those areas that had recently gone (or were about to go) through modernisation. It also allowed us to engage in discussions with curriculum teams within SQA to consider how they could engage in developing eAssessment within their area.

CPD Benefits

On a positive note, participants in the evaluations believed that the project had value in terms of Continuing Professional Development benefits – while the SOLAR project is primarily about creating assessments, a key achievement has been enabling colleagues to work together in ways that they may not ordinarily have been able to. In particular, the project gave them the opportunity to discuss the assessment of their subject area with a group of their peers. Many of the authors who participated in the evaluations also thought that involvement in the project had had a positive impact on their own professional practice. From initially seeing the technology as the limiting factor, they had now moved to seeing the limitations of the qualifications frameworks as one of the major barriers. This has led some of the authors to further develop their skills and qualifications in this area by looking to study towards the *Diploma in eAssessment*.

Dissemination and Support

As assessments are completed, peer reviewed, moderated, and quality-assured, the project moves from development into delivery, dissemination, and support (although more assessments will continue to be rolled out as they become available).

Across Scotland over 80% of FE colleges have received training in the use of the administration system to support the delivery of the assessments from the project. In most colleges this has been a single individual, although in a few colleges three or four staff have been trained. This is has been in response to their own plans for devolved administration of the assessments within different

areas within the college. The role of the centre administrator for the SOLAR project varies from college to college, depending on staffing levels, curriculum requirements, and internal structure. The majority of those attending the training had a role in supporting general eLearning either within an individual department (usually Computing) or the college as a whole. In colleges where there was no departmental support available, this role was taken on by someone within student records or the exam office.

Although we provide advice and support to colleges on appropriate procedures to be used within their centre, individual colleges are able to produce their own procedures on the scheduling and delivery of the assessments. Scottish further education colleges have devolved responsibility to develop and maintain their own quality assurance procedures. Therefore, the responsibility for a centre in delivering these assessments is no different than if they were delivering a traditional, paper-based assessment. Both are subject to the college's own quality assurance procedures, which in turn have been approved by the education inspectorate, HMIe.

Many colleges have already considered how eAssessment delivery might impact upon their assessment procedures and these have been implemented to ensure the effective delivery of eAssessment.

It is important to note that, even at this stage, SOLAR is still a work in progress. The iterative process used in the development of assessments will continue, and feedback obtained from users – both students and staff – will influence the modification and updating of the assessments.

User Experiences

To support students undertaking the assessments, a flash-based tutorial has been produced which enables the learner to practise how to navigate and use the delivery system. Not only is this tutorial available on the website, it is also available within the delivery system so a student may use it, with no loss in assessment time, to practise before undertaking the assessment.

In terms of our evaluation of the success and effectiveness of the project, learner feedback will provide a valuable addition to the views and experiences of the authors and moderators. As such, processes have been put in place that mean we can reflect and respond to feedback. We have urged centres to encourage their students to complete a post-assessment questionnaire. This evaluation survey of opinion and comment is available online and in paper format, and targets issues such as whether students felt they were adequately prepared for their assessment, whether they have had any previous experiences of eAssessment, and how easy they found the assessment system to use. It also evaluates learners' views on the contents of the assessment, the system of immediate feedback, and whether their preference is for traditional, paper-based assessment or online, automatically-marked assessment.

User feedback is in its very early stages; we have some qualitative feedback, but we await more data for quantitative evidence. However, output so far from evaluations suggests positive experiences from both staff and learners. Learners like the immediate feedback of score and the interactive nature of the assessment delivery system. However, the delivery system we use does not provide question by question feedback directly to the learner at the end of the assessment. This is because we see this feature as being related to the teaching and learning taking place and, hence, this should be managed by the tutor. Therefore, if feedback is required for learners then the tutor may go into the web-based management system and check individual student answers. They can then use their own professional judgement to provide appropriate remediation and support before the student undertakes another version. We consider this feature to be an important mechanism to support tutors' wider teaching strategy. In early evaluations, a few students did comment that they would like full feedback for each question, particularly ones that had been answered incorrectly.

More detailed user evaluation should help us identify and address issues as the project progresses, and this has already been identified as a key activity for our work as the project progresses into 2007.

Identifying Development Priorities

Development priorities are impacted by different factors such as: uptake of qualifications; acceptance of eAssessment within the sector; new technology; changes in curricular requirements (unit specifications); and similar activity within the assessment field. The output of user evaluation reflects these factors. Positive evaluative feedback from students in a particular area might indicate increased demand for eAssessment, so encouraging the development of eAssessments in that area of the curriculum.

The initial feedback from staff involved at different stages of the project has already had an effect on the identification of the priorities for the next stage of activity. The need to provide enhanced support to centres using the assessments and the need to deliver develop better links between SOLAR and other eLearning and eAssessment projects have been recurring themes in this feedback. Further development of the assessment management and reporting system to reflect the requirements of centres is also planned during 2007. Furthermore, in response to demands from the sector, we will be developing assessments within Automotive Engineering and Horticulture during the next phase of the project.

The processes and procedures used in these developments are sometimes just as important as the suitability of a curriculum area and the skills of the authors in development. It is essential that the development of eAssessments is supported by effective management and quality assurance. We believe that the effectiveness of these processes can only be ensured by engaging participants in the decision-making process.

Future Steps

Project activity will move from being mainly assessment development focused to address four particular areas which we have identified as being crucial to future direction.

- Supporting centres in delivering the assessments, with a greater emphasis on evaluation of staff and student experiences. This evaluation will have a significant impact on the nature of ongoing development within and outwith the project.

- Working closely with other eLearning and eAssessment projects to engage in dissemination, evaluation, training, and promotion, while at the same time providing a co-ordinated approach to supporting the blended learning agenda.

- Making the case for continued development of eAssessment on the grounds of sound pedagogy, while evaluating its ability to provide a reliable and robust method of assessment of knowledge and skills at different learning levels

- Continued development of eAssessments across a limited number of new subject areas. This will involve the project in work within new areas, such as Automotive Engineering, Sports, and Horticulture. The approach taken here will continue to reflect the evaluation and feedback from the existing development teams.

Although the project is limited in scale to a selected range of curriculum areas, it has the potential to make a significant impact in the Scottish Qualifications Authority, and we expect the work done in this area will continue to have a growing effect on the nature of qualifications and how they are assessed in years to come.

GENERALISE NOT SPECIALISE: DESIGN IMPLICATIONS FOR A NATIONAL ASSESSMENT BANK

Rod Johnson and Sandra Johnson

Generalise not specialise: design implications for a national assessment bank

Rod Johnson and Sandra Johnson
Assessment Europe
mail@assessment.eu.com

Abstract

Within the framework of the Assessment is for Learning (AifL) programme[1], two systems of national assessment are currently operating in Scottish schools: on-demand 5-14 National Assessments and the sample-based Scottish Survey of Achievement. This paper will discuss issues surrounding the design of an assessment bank intended to support both systems.[2] It focuses in particular on the considerations underlying decisions about the structure of the shared materials database, the complex definition of an "item" that had to be adopted in order to accommodate a wide range of assessment types, the overall architecture of the wider information system, with its component databases (one being the bank) and information management subsystems, and the tensions arising from the need to accommodate the requirements of different systems of assessment while avoiding the dangers involved in data repetition and redundancy.

Introduction

Since the autumn of 2003, primary and lower secondary teachers in Scotland have benefited from online access to 'national assessments': these are tests which they can use on a voluntary basis to confirm their judgments about their pupils' levels of attainment in reading, writing and mathematics (for level descriptions see the relevant 5-14 curriculum guidelines: SOED 1991a for English language, 1991b and 1999 for mathematics)[3]. Schools make requests for assessments through a web interface[4], identifying their needs in terms of subject and level. A school might, for example, request a 'Level B' assessment in mathematics or a 'Level D' assessment in reading.

In reading, an assessment comprises two different tasks, where a task consists of a source text plus multiple associated test questions (20-30,

[1] www.ltscotland.org.uk/assess/
[2] There is, of course, an important third area of pupil assessment on a national scale in Scotland: external examinations, run by the Scottish Qualifications Authority (SQA). While the work discussed here is not mandated to cover application in the area of external examinations, we have tried as far as possible to keep the design we propose sufficiently flexible to accommodate this major category of high-stakes, pupil-based assessment.
[3] It is likely that at some point national assessments will be extended to include science and social subjects.
[4] www.aifl-na.net

depending on level). In mathematics, assessments comprise two loosely parallel 'booklets', each comprising 20-30 'atomistic' test items, all at the same level but spanning the mathematics curriculum at that level. Reading task pairs are selected at random from within a pool of appropriate assessment materials in response to individual requests. Mathematics booklets are created using domain sampling, i.e. random selections of items are drawn from within the materials store, following a test specification that dictates overall item numbers as well as imposing some constraints on content coverage.

The Scottish Survey of Achievement (SSA)[5], on the other hand, is a programme of annual sample-based surveys of pupil attainment at selected stages in primary and early secondary education. The SSA, launched in 2005, evolved from the Assessment of Achievement Programme (AAP), which was introduced in the mid-1980s and ran until 2004. The distinctive feature of the SSA is that pupil attainment is reported by individual local authorities as well as nationally, whereas the AAP reported only nationally. Attainment is currently reported for four subject areas, assessed on a 4-year rolling cycle – English language, mathematics, science and social subjects; core skills feature every year (reading, writing, numeracy, ICT, problem solving and working with others). In certain cases, domain sampling is employed to select items and to create tests for survey use. In all subjects, items are randomly allocated to pupils using multiple matrix sampling. Pupils' attainments are typically reported in terms of proportions attaining given 5-14 levels, using the same level descriptions as national assessments and the same decision criteria.

Both national assessments and the SSA assess pupils' attainments in essentially the same way, using the same kinds of assessment materials; indeed, materials used in the SSA are available post-survey for use in national assessments, and materials developed independently for use in national assessments are available also for survey use. Unsurprisingly, the decision was taken to maintain a shared resource of assessment materials, which we can call the 'assessment bank'. The assessment materials already in the bank[6], and others soon to be incorporated, are quite varied in nature, ranging from typical objective and short-answer forms to structured questions and themed item sets (e.g. reading tasks). Practical assessments of various types feature in the attainment surveys in all subject areas, and at some point these, too, will be banked.

But while the attainment surveys and national assessments draw largely on the same basic stock of assessment materials, the needs and aims of the two programmes are essentially different. One programme is pupil-based, and intended to provide teachers with information about their pupils to use when evaluating individual progress and determining next steps in learning; the other is cohort-based, where individual pupil assessment is subordinate to the gathering of information about the performance of the education system as a whole. These differences have significant implications for the structure and

[5] www.ltscotland.org.uk/assess/of/ssa/

[6] The Scottish Qualifications Authority is responsible for developing and maintaining bank content.

content of the assessment bank. In particular, it is important to maintain a perspective on the bank not as an isolated entity, but as one component of an evolving, larger and more complex information handling system targeted on assessment applications. We discuss the wider, dynamic context below, but first we need to consider the range of static information stored in the bank itself.

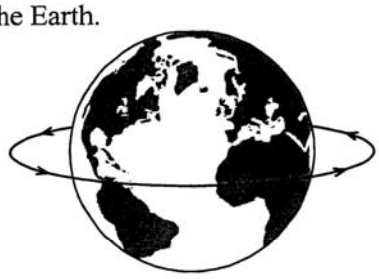
The assessment bank

Our design for the 5-14 national assessment bank, and for the SSA and national assessment information management systems, is based on several years' experience during the late 1990s/early 2000s, recovering historic AAP assessment materials and associated performance data and developing a prototype information management system for the programme (Johnson & Johnson, 2002 and 2003).

We came very early to the realisation that there was no simple organisational structure that would readily handle the wide variety of assessment materials used in the attainment surveys, in multiple subjects across a broad range of pupil ages. In particular, it was never going to be acceptable to design a banking system constrained to accept only objective format items[7], of which Figure 1 reproduces a typical example.

Source: AAP 2005a, Chapter 2, page 11

This is a picture of the Earth.



The Earth spins round and round. The arrows show it spinning.

How long does it take the Earth to spin round once?

Tick (✓) the correct answer.

1 hour ☐ 1 day ☐ 1 week ☐ 1 year ☐

Figure 1: A multiple-choice science item

[7] The developers of the national assessment bank in its present form failed to fully appreciate this, with the consequence that the bank now needs to be re-structured to accommodate the greater variety that was always present in the set of assessment materials used in past and current surveys.

Reading tasks offer the most extreme examples of assessment materials that do not fit the objective format item mould. These comprise a source text, or 'passage', followed by a relatively large set of questions, or items, grouped into 'sections', usually on the basis of a common format (see Figure 2).

Source: SSA 2006, Technical Annex, Section C, page C4

Stimulus text

In this example, *Attila the Hen*, a 420-word passage recalls events at Sunnycluck Farm just after all the hens have made their escape. Attila realises that the other hens are looking to her to lead them.

Section A: 10 multiple-choice questions

Section B

Arrange these sentences in the right order by putting the correct letters into the boxes below. The first one is done for you.*

A. The dogs hear Attila's squawk.

B. The hens return to the farmyard.

C. Attila leads the escape.*

D. The hens are upset by Attila's orders.

E. The men try to round up the hens.

F. A group of hens gather together.

*	1	2	3	4	5
C					

Section C

Here is a summary of part of the story after the farmyard battle. Fill each gap with **one or more words**. You may use words from the story or your own words.

Attila watched as the men returned to the _____ (1).

She decided to find out if _____ (2) had survived.

Taking a _____ (3) she _____ (4).

Soon she had assembled _____ (5).

Attila realised that her _____ (6), the old hen,

was _____ (7).

She decided she would have to _____ (8) the other hens

by herself because they _____ (9) her.

Figure 2: A typical reading task structure (abridged)

To handle this kind of assessment, we consider the typical unit of presentation in an assessment to be a *task*, perhaps with *subtasks*, containing *items*. An item is the smallest element of assessment with which we can associate a *score*.

It is, however, not always clear what exactly is the precise decomposition of a task into its constituent items. For example, does the task in Figure 3 contain a single item? Or three items? Or six?

Source: AAP 2005a, Chapter 2, page 10

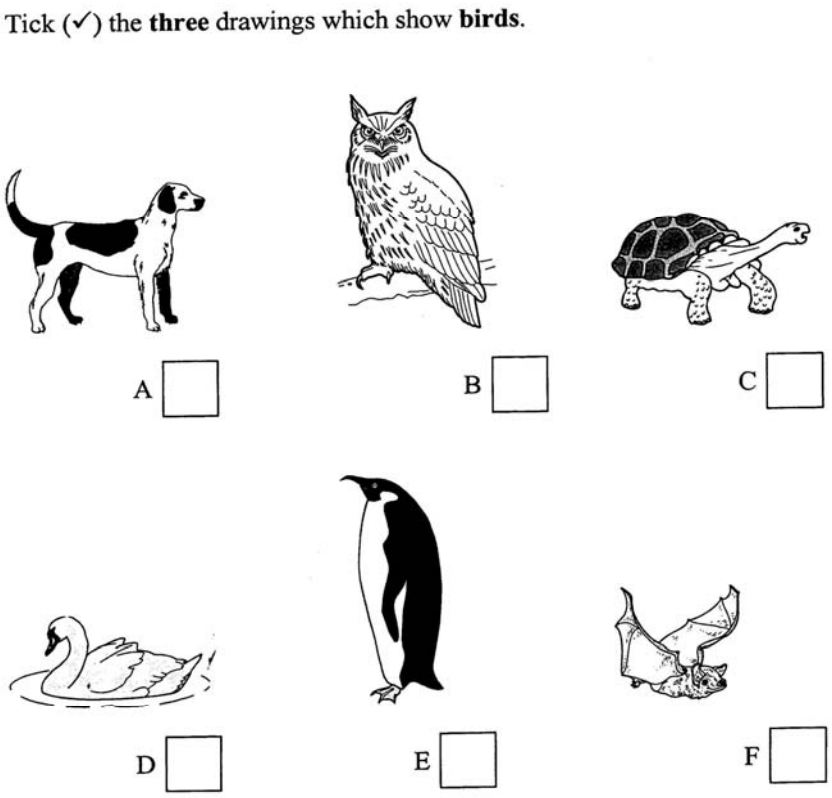


Figure 3: A science task comprising 'subitems'

Examples like this suggest that it may be useful to introduce a level of description below that of an item – a kind of atom to the item's molecule. Our design includes a notion of *subitem*, an element which can be associated with a pupil response, but which can only participate meaningfully in scoring its containing item when taken in conjunction with its fellow subitems.

Note that a response is not the same as a mark or score. The response is, ideally, the transcription of a subject's actual answer to the (sub)item, perhaps mapped to one or more of a finite set of possible responses, not all of which need to be correct; in the less ideal, but frequent, case where only information supplied by a marker is recorded, the response is just the marker-supplied information, again possibly mapped into a prescribed finite set. The (sub)item

mark is a binary quantity, representing the dichotomy correct/incorrect, derived by rule from the response: this is what we call the subitem *mark*. Where marker information only is recorded as the response, the relation between a response and its mark is just identity.

We define an item *score*, on the other hand, as a function of a set of subitem marks together with a rule for computing a composite numerical value from the responses, called a *mark scheme*. Subtasks and tasks also can have scores, usually computed by relatively trivial mark schemes (simple summation, for example).

The complete structure which we have currently implemented to handle the storage of the assessment materials is outlined in Table 1.

Task:	a set of questions, grouped into one or more *sections* or *subtasks*, normally based on a shared stimulus (text, picture, video clip …)
Subtask:	a collection of one or more *items*, based on the same stimulus material and usually, though not necessarily, sharing other common properties such as format, theme, level of difficulty; from the point of view of presentation subtasks are often labelled as *sections*; a subtask is characteristically the smallest unit of assessment whose external form can be independently stored
Item:	normally the smallest element of assessment which can be *scored*, though computation of the item score may involve consideration of *responses* to several constituent, usually interdependent, *subitems*
Subitem:	the smallest element of assessment for which a response can be recorded.

Table 1: A generalised ontology for storing assessment materials

In many types of assessment, an item and the corresponding task are expected to be equivalent (i.e. the task, subtask and item each contain just one component), the item has just one subitem, and all associated mark schemes are trivial, as is the case with orthodox multiple-choice items. Figure 1 above is an example, as is Figure 4 below, of what we often call a 'single-item' task.

Source: AAP 2005b, Chapter 2, page 9

Solve the following equation.

$$3(x-2)+7x=24$$

Answer: $x=$_____

Figure 4: A short answer mathematics item

This kind of item, however, in many assessment contexts, is very much a special case, and not the norm.

Consider the example in Figure 3 above: we treat this as a task having a single constituent item (and hence *a fortiori* just one subtask), the item having six subitems. The score for the item is a function of the set of responses to all six subitems (some of which may be blank).

In other examples, such as that shown in Figure 5, there is a clear composite structure, with a 'task' comprising (a single subtask with) two or more 'items'. Here the first item is a short answer question, while the second invites a more extended open-ended response. While the two items focus on the same general concept of force, they are in fact independent. Each could be presented quite separately, even without the introductory sentence and diagram (with a minor word change to the second item). But as they stand, from a presentational viewpoint there is little to be gained by storing them separately.

Source: AAP 2005a, Chapter 2, page 13

A spring balance can be used to measure the force exerted by something.

a. What is the unit of measurement of force?

_____

b. The box has a mass of 1kg but the scale on the instrument reads 10.

Why is this?

_____

_____

_____

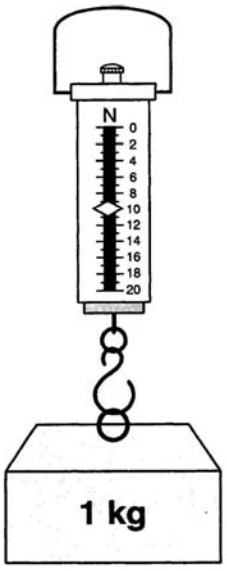


Figure 5: A composite 2-item science task

Figure 6 overviews a mathematical literacy task. This is a task typical of its kind, comprising a series of items based on a common stimulus. While independent in the sense that a correct answer to one item would not increase the chances of a correct answer to any other, the items in this case could not be presented separately from the others without reproducing all or the relevant part of the stimulus materials.

An interesting case of a multi-item reading subtask is that of a summary completion exercise (see Figure 2). Pupils are invited to fill gaps in a short summary of a longer text, implicitly reproducing the sense of the original whilst maintaining grammatical integrity. Here, each 'gap' is essentially a separate test item, but it would not be possible to present any item separately from the rest.

Source: AAP 2005b, Chapter 2, page 12

'Crime Survey'

The source material for this task comprises eight pie charts, illustrating the results of a survey into people's experience of crime. Each pie chart shows the proportion of individuals in the crime survey who answered in particular ways to questions such as "Have you, or another member of your immediate family, been a victim of crime in the last five years?" (response options: 'yes, self'; 'yes, other family member'; 'no'). Pupils are asked 12 questions, all requiring them to read information from one or other of the charts: five are short-response items, including "What percentage of people surveyed had **personally** had a crime committed against them in the last 5 years?" and seven are multiple-choice items.

Figure 6: Overview of a multi-item mathematical literacy task

Tasks, subtasks (to a lesser extent) and items have associated descriptive metadata, which we do not have space to go into here. Resource materials also have a set of associated descriptive metadata; where possible the resources themselves are incorporated into the bank.

A distributed information system

We said earlier that the assessment bank should be seen as just one component of a more complex architecture. To see why this would be so, recall that the materials in the bank should be directly available for use in at least two distinct contexts: the national assessments and the SSA.

While the system of national assessments is at present essentially a one-way communication system, in the medium term it is planned to develop the system further, in particular by facilitating 2-way communication, for example to receive pupil performance data from the schools, to provide feedback in the form of comparisons of class/school performance with national results (using SSA data), to allow pupils to take tests on-line, and/or to offer automatic marking to those teachers who request it.

For its part, administration of the SSA involves sampling from national school and pupil populations, communicating with authorities, schools and other

organisations, receiving, validating and processing pupil response data, carrying out automatic marking of item responses, producing a standard set of summative attainment reports, and keeping records of all of this activity as well as archiving response data at a detailed level for later retrieval for a variety of purposes.

It is evident that any attempt to incorporate one or the other of these functionalities directly into the bank design could risk prejudicing its utility for the other application. Moreover, the two functional descriptions above effectively describe the basic requirements for a pair of information management systems (IMS), respectively oriented towards the administration of on-demand test delivery and national system evaluation. These two observations together motivate our design for the union of the national assessments and the SSA into a distributed information system, based on a conceptual and organisational separation of the static, intrinsic characteristics of the materials themselves (the assessment bank proper) from dynamic, application-generated information (usage and performance data, *inter alia*, contained in dedicated information management subsystems).

A second shared resource, discussion of which is beyond the scope of this paper, is the set of externally maintained information about schools, information that is essential to the survey programme for sampling, distribution and analysis purposes and to the national assessments programme for the authentication and monitoring of requests from schools for assessments.

Figure 7 illustrates schematically the overall architecture of the system.

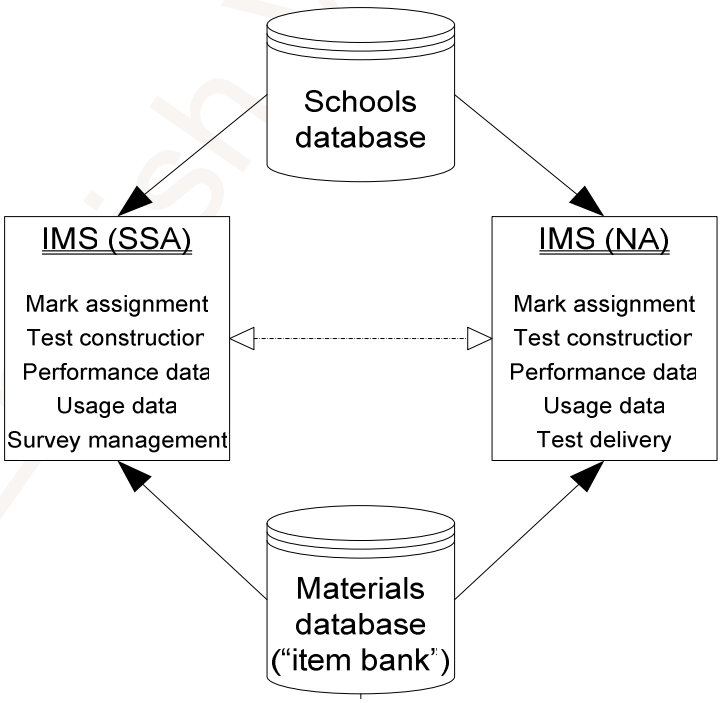


Figure 7: Distributed information system architecture

Note that information flows essentially in one direction from application-neutral databases to application-dependent IMS. At the same time, we would prefer

to minimise traffic between one IMS and the other, as symbolised in the diagram by the dotted line connecting the two, so as to allow as far as possible development to proceed independently.

As an example of the tension that can arise out of these constraints, consider the case where the developers of the national assessment IMS might choose to use item facility as part of a strategy for determining dynamically the balance of items in a test. Given that the same items are potentially used in SSA surveys, they would like to use relevant SSA performance data to produce the required facility estimates. So now the question arises as to where such data should reside, with the obvious temptation to store them directly alongside the items within the assessment bank. We are not in favour of this approach, for several reasons:

- such facility estimates are subject to dynamic change, as opposed to the stable, static information typically housed in the database;

- item facility is population-dependent, which means that facilities computed in one testing context might not be relevant for use in another;

- in any case, good data management systems design suggests that, *ceteris paribus*, values that can be readily computed from existing data should not be stored independently; if we follow this logic, we would have to consider storing the raw SSA responses themselves in the assessment bank; and if we store the SSA responses, why not the national assessments responses too?

- allowing the SSA IMS to deposit its results inside the assessment bank takes away control of the content of the bank from its own administrators.

On the basis of this kind of argumentation, we have had to conclude that a limited measure of interaction between constituent IMS has to be allowed, in order to maintain the autonomy and integrity of the assessment bank. Even so, we attempt to enforce the principle that all such interaction should always be subject to careful, bilateral negotiation.

Indeed, the question arises generally: what should form the content of the bank and what should more appropriately be located within the two dedicated IMS?

As we have just argued, we believe that item performance data appropriately belongs within the respective IMS, ideally in the raw form of pupils' qualitative responses to (sub)items, and not in the form of summative scores (these can at any time be generated on demand from the detailed response data). Similarly, usage statistics, those dynamic tracking statistics that monitor the use of individual items, tasks and tests, should also reside within each applications-specific IMS.

On the other hand, we have through experience come to the conclusion that, in addition to the assessment materials and associated descriptive metadata, the bank itself should hold a set of response options for each (sub)item, where

a set might comprise a single 'right answer', multiple alternative right answers or a series of right and wrong answers (which could have diagnostic value).

Mark allocations, however, and the construction of marking algorithms and mark schemes, should, in our view, more properly be held in some appropriate form within the applications-specific IMS. This is because mark allocations, even in the case of clearly identifiable individual items, can vary, depending on the purposes of the assessment application, the subject being assessed, the ages of the pupils/students being assessed, and the predilection of the assessors involved.

Test generation, too, should in our view reside within the IMS and not within the assessment bank. Again, this is because different applications can, and in our case do, use the same assessment materials to create quite different types of test or to create similar tests packaged differently.

In the national assessments, for example, mathematics tests are produced by drawing random samples of mathematics items from within the assessment bank, to provide an agreed representation of the curriculum, but with all the items at the same 5-14 level. In the recent 2005 SSA, numeracy tests were created in a similar way, but this time each test included items at three different 5-14 levels, with a randomised item ordering within the test itself. In both application areas the test generation algorithm might also change over time, another reason to keep this facility within each applications-specific IMS.

Finally, of course, each IMS is designed to deal with all the administration and transactions which characterise its particular application. In the case of the SSA, for example, the IMS should be expected to handle, *inter alia*, the generation of form letters to authorities and schools involved in the survey, specialised analyses of the results, and production of routine tables and reports.

In conclusion

The overall picture is one of some complexity, far greater than can be accommodated by the homogeneous, monolithic structure we might expect to find in conventional 'item banks', of the type implied in the oft-quoted definition (Sclater & McDonald 2004):

> "A collection of items for a particular assessment, subject or educational sector, classified by metadata which facilitates searching and automated test creation."

After several years of maintaining an archive of AAP survey materials and results, when faced with the challenge of designing an integrated resource which would serve adequately both the AAP's successor, the SSA, and a system of nationally available on-demand assessments, we concluded that the appropriate architecture was not an item bank as generally understood, but a distributed information system.

The system draws on the materials stored in an assessment bank as well as on shared information about schools and pupils. The bank is designed to

accommodate the wide variety of items and tasks that continue to be favoured by test developers in the different subject areas, to address assessment requirements at different levels in the education system, and to serve the specific needs of the different application domains. These are the two senses in which we have generalisation.

At the same time, we should be extremely careful not to bias the bank by imposing structures and behaviours which are largely the preserve of one application area, perhaps even to the extent of being in conflict with the needs of the other. Finally, we strive in our design, insofar as we are able, not to prejudice future extension of the bank to other, distinct applications (certification and selection, for example). In designing and developing such a complex artefact, tensions are bound to arise between, on the one hand, the conflicting requirements of different assessment needs within the system, and, on the other, the desire to maintain as high a degree as possible of application neutrality in the bank. Wherever possible we have tried to use principles of sound system design to resolve such tensions.

References

AAP (2005a). *The Sixth AAP Survey of Science (2003)*. Edinburgh: Scottish Executive Education Department.

AAP (2005b). *The Seventh AAP Survey of Mathematics (2004)*. Edinburgh: Scottish Executive Education Department.

Johnson, S. & R. (2002). *An architecture for the 5-14 Assessment Information System*. Internal report to the Scottish Executive Education Department, December 2002.

Johnson, S. & R. (2003). *A suggested basic structure for the National Assessment Bank*. Internal report to the Scottish Executive Education Department, April 2003.

Sclater, N. (2004), ed. *Item Banks Infrastructure Study (IBIS)*. www.toia.ac.uk/ibis.

Sclater, N., McDonald, M. (2004). 'Developing a national item bank', *Proceedings of the Eighth International Computer Assisted Assessment Conference*, Loughborough University, July 2004; cited in Niall Sclater (2004), ed.

SOED (1991a). *National Guidelines: English Language 5-14*. Edinburgh: Scottish Office Education Department.

SOED (1991b). *National Guidelines: Mathematics 5-14.* Edinburgh: Scottish Office Education Department.

SOEID (1999). *National Guidelines: Mathematics 5-14 Level F*. Edinburgh: Scottish Office Education and Industry Department.

SSA (2006). *The First SSA Survey of English Language and Core Skills. Technical Annex*. Edinburgh: Scottish Executive Education Department (in press).

MAPLE T.A.
A SPRINGBOARD FOR WEB-BASED TESTING AND ASSESSMENT

Samir Khan

Maple T.A.
A Springboard for Web-based Testing and Assessment

Samir Khan
Adept Scientific
Amor Way
Letchworth
SG6 1ZA
01462 489126
Samir.Khan@adeptscience.co.uk
www.maple.adeptscience.co.uk

Abstract

Maplesoft products have helped institutions offer more progressive instruction to students in mathematics, engineering, and science for over twenty years. Their software tools have made flexible and intelligent mathematical computing an essential component of modern education. They are a major supplier of maths software tools to educational and commercial institutions in the UK, Europe and North America.

Maple T.A. is a web-based tool for testing and assessment of any mathematical concepts and builds on this rich history by combining Maplesoft's extensive education experience with the best in online technology. Maple T.A. supports complex, free-form entry of equations and intelligent evaluation of responses that can test for algebraic equivalence, making this system ideal for mathematics, science, or any course that requires mathematics. It is ideal for placement testing, homework delivery, drill and practice, exam questions and assignments, high stakes testing, standards and gateway testing, and "just in time" teaching.

Maple T.A. has the full power of the advanced mathematical software Maple behind it. Maple has the ability to represent and solve problems in calculus, linear algebra, abstract algebra, vector calculus, statistics, number theory, group theory, and more. The Maple engine can determine mathematical equivalences, and automatically grade the student response appropriately. Like a human teacher, Maple T.A. will detect when the response is equivalent to the programmed answer, instead of doing a mindless simple comparison. Maple T.A. can display MathML, so all equations look the same as they do in your textbook Students can use palettes and a math expression editor with free-form input, so they can enter their responses in the same way they would write them down. To simplify entry, a 1-D graphing calculator is available. The student then has the option of previewing the equivalent 2-D expression before submitting their response. Maple T.A. includes built-in unit support for

many common types of measurement. Equivalent answers will be graded correctly. For more obscure measurement systems, Maple T.A. can be programmed to evaluate the problem. Maple has an extremely thorough coverage of units. Maple T.A. can be used directly from inside your Blackboard classes.

This presentation gives an overview of Maple T.A. The general features and philosophy will be explored, followed a practical demonstration of the product, filtered through both the student and teacher experience.

Adept Scientific are the full-service partners for Maplesoft in the UK and Northern Europe, and have delivered technical computing solutions to mathematicians, scientists and engineers for over twenty years.

QUICKTRI AND THE INTELLIGENT SHELL SYSTEM (ISS) FROM INNOVATION 4 LEARNING - BUILDING ON PRACTICAL EXPERIENCE

Don Mackenzie & Matthew Stanwell

QuickTrI and Intelligent Shell System (ISS) from Innovation 4 Learning. Building on Practical Experience

Don Mackenzie & Matthew Stanwell
Innovation 4 Learning
Business Development Unit
University of Derby
DE22 1GB
D.Mackenzie@derby.ac.uk

Abstract

The increasing amount of commercial work being undertaken by the Centre for Interactive Assessment Development (CIAD) and Interactive Media Unit (IMU) at the University of Derby has promoted the separation of commercial activities from the day-to-day academic work of those departments.

This has resulted in the development of a new commercial e-business division within the University, Innovation 4 Learning (i4L). Here a core team derived from both CIAD and IMU have joined with commercial professionals to focus on bespoke applications for clients in the corporate, health and educational sectors initially together with the development of two core products.

The Intelligent Shell System (ISS) enables anyone to rapidly create and maintain a professional-looking and engaging, web-based e-learning site that can incorporate documents, graphics, video sequences, Flash™ objects and quizzes, without specialist technical skills. Menus are extensible and fully configurable.

QuickTrI is a development of the longstanding TRIADSystem with an intuitive WYSIWYG design interface for the rapid development of highly interactive assessments that harness the full power of the TRIADS delivery engine together with the sophisticated scoring capability and configuration control of the current TRIADS question templates. The system will be ideally suited to highly visual disciplines in allowing extremely flexible screen designs that that may incorporate overlaid and intimately embedded text, video and graphics.

QuickTrI assessments may be called from and send results to ISS sites.

QuickTrI is due to be released early in 2007 but delegates will have the opportunity to comment upon and contribute to the work in progress on the design interface both in the presentation and on the i4L stand in the exhibition area. Assessments created with the system may be stand-alone, CD-ROM, LAN, Intranet or Internet delivered.

LEARNING FROM ASSESSMENT: EVALUATING THE BENEFITS OF DALI (DIAGNOSTIC ASSESSMENT LEARNING INTERFACE)

Hershbinder Mann and Guinevere Glasfurd-Brown

Learning from Assessment: Evaluating the Benefits of DALI (Diagnostic Assessment Learning Interface)

Hershbinder Mann and Guinevere Glasfurd-Brown
University of Essex,
hmann@essex.ac.uk
guin@essex.ac.uk
01206 872026

Introduction

The DALI Project is a Teaching and Learning Innovation Fund (TALIF) pilot study at the University of Essex. DALI (Diagnostic Assessment Learning Interface) is an add-on to QuestionMark Perception (1), which enables students to see how well they've performed in each topic (a group of questions). It provides learners and instructors with multiple ways to view assessment information in order to gauge progress in specific topic areas.

The DALI interface:

- allows a student to select the score level that they wish to achieve;
- highlights the last topic-based feedback provided to the learner;
- provides statistics associated with any selected topic; and
- displays topic descriptions as learning objectives.

The paper discusses how the project aimed to understand better the use of online assessment with a particular focus on the student experience. Data were collected on matters such as what the students perceive to be the merits and demerits of online assessment, and what motivated or discouraged them from using it. These issues are particularly salient: the National Student Survey has shown assessment and feedback to be a particular concern for students. As the market for e-delivery expands more thought needs to be given to how students learn with e-learning, and the ways in which this should inform the design of e-learning activity, including e-assessment.

Overview of the DALI Project

Interest in online assessment has grown rapidly at the University of Essex and is a key area for development nationally. The University has developed a student portal and is developing personalised assessment resources as part of the FDTL5 SPRInTA Project (2004-06, http://www.essex.ac.uk/sprinta). Previous TALIF projects have developed online assessment to meet specific

departmental needs and concerns, including using formative assessment to support large group teaching and to support threshold testing.

The project also aligns with developments at a national level: Assessment and personalised learning opportunities are included in e-learning strategies from the DfES and HEFCE. In January 2005, the Qualifications and Curriculum Authority published a 'Futures'(2) paper: which also identifies e-assessment as a means of supporting personalised learning, which they argue is a key area for curriculum development in the next 10-15 years.

Alongside this, UK HE funding bodies have articulated the need for research to be undertaken into student uses and experiences of e-learning. The JISC has funded a strand to the e-pedagogy programme (2004-07) entitled, 'Understanding my learning', which focuses on the learner perspective on the role of ICT in learning: '*Learners have different priorities, preferences and approaches to learning, and different requirements for support. The learning environment needs to reflect these differences. Understanding how different learners experience the tasks, resources and services offered to them is an important precursor to developing effectively personalised systems' (3)*. The DALI Project sought to assess the student experiences and use of online assessment and will incorporate these findings within a good practice guide on online formative assessment to be made available to staff from October 2006.

The DALI Project

The DALI Project aimed to build upon online assessment experience in three departments at the University of Essex. The departments of Accountancy and Financial Management, Electronic Systems Engineering, and Biological Sciences have used QMP in recent years to generate a large number of question items and re-usable databases to support formative testing. The DALI Project Officer worked with academic staff in these departments to add topic-level learning objectives and feedback into the weekly assessments for three courses. These formative assessments were then made available to students in the Autumn term 2005, with evaluation on staff and students perspectives on online assessment in the Spring term 2006.

The DALI Project evaluated both student and staff perspectives on online assessment through the use of a student questionnaire, a student focus group, and a focus group session involving both staff and students (held in the University's i-LAB). The initial student survey was conducted electronically and the findings from the survey formed the basis of an assessment 'think tank', a focus group session which discussed assessment and feedback with students.

When thinking about assessment, 89% of students surveyed either strongly agreed or agreed that being able to identify their strengths and weaknesses was important to them, and 92% strongly agreed or agreed that is was important to be able to track their own progress. 60% of students either strongly agreed or agreed that they used the formative tests for feedback, however 60% of students also strongly agreed or agreed that the feedback could be improved.

An iLAB meeting with staff and students, in January 2006, was particularly

productive, and used the environment to discuss perspectives on assessment, focusing around a discussion of the possible strengths, weaknesses, opportunities and threats of formative online assessment:

Strengths and Opportunities – (Students)

1. Enables me to identify strengths and weaknesses in a subject
2. Provides immediate feedback
3. Great for revision
4. Available to use throughout the year
5. Could be used to personalise learning
6. Builds confidence
7. It might highlight strengths students weren't aware of
8. Could be expanded into non-academic areas
9. Useful for personal development – control own learning

Strengths and opportunities – (Staff)

1. Once running, low maintenance cost
2. Feedback can include new material that's not explicit in the course notes
3. Feedback can included URLs and point to relevant section in lecture material
4. Easily extended and updated
5. An effective means of integrating learning and assessment
6. Reduce workload
7. Feedback from students (use free-text question at end of test for comments from students)

Weaknesses and threats – (Students)

1. Failing tests can be demoralising
2. Not enough feedback is personalised
3. Access issues
4. Poor report display
5. Online testing not used to full potential (too many MCQs)
6. Decrease in student-staff contact time

7. Possible that students could become expert in tests rather than learning

8. Too assessment oriented – constantly tested. More than one way to learn

9. Increase workload and pressure

Weaknesses and threats – (Staff)

1. Up front cost in time (skills)

2. Authoring good feedback can be difficult and is time-consuming to generate

3. “Distance” - lack of direct contact with students

4. Administrative burden

5. How to engage students? Substitute

6. Cost; continuing support

7. Support logistics

Students' perspectives identified a number of clear benefits and learning advantages to online formative assessment: For example, the students commented on the usefulness of the tests for revision; the ability to identify strengths and weaknesses; and the fact that QMP is not used to its full potential, particularly in the area of personalised feedback (i.e. feedback which draws together data from across an assessment to create a unique learning profile, rather than generic feedback by score band).

Staff concerns often centred on time, workload and support issues. For example, staff worried about the resources needed for in the development of question banks and assessments and the energy required to author effective feedback. Staff also commented positively on the possibilities to extend and update the tests.

The question of feedback was one of the most discussed topics with differences of opinion in how detailed this feedback should be. Students favoured comprehensive feedback that included explanations of why the correct answer is right, and why the other answers might be wrong. Feedback that stated simply whether a response was correct or not was not deemed to be as useful. The staff who participated in the session, were cautious of explaining the right and wrong answers for every question. There was the suggestion that the feedback should instead point the student to sources (for example a URL to assigned reading, lecture notes etc.) that could be of use in answering a particular question. It was felt that this would not only address student concerns but also enhance independent student learning.

Conclusion

The DALI project was a small pilot, which ran across three courses for the period of one term. The findings from the survey, the focus group and the iLAB session have since been incorporated into a staff handbook on online formative assessment, to enable students and staff to derive optimal benefits from the use of online assessment for formative testing.

The project experienced some problems with the DALI interface during the pilot, which are currently under investigation. However, the project was instructive in highlighting the ways in which students approach formative online assessment, and the potential they saw for personalised learning. The University is currently working towards developing an 'intelligent' assessment environment to support interprofessional learning which is capable of generating personalised learning profiles as point for reflection on learning.

References

(1) http://www.questionmark.co.uk/uk/perception/dali.htm

(2) http://www.qca.org.uk/downloads/futures_meeting_the_challenge.pdf

(3) http://www.jisc.ac.uk/index.cfm?name=elearning_pedagogy

FORMATIVE ASSESSMENT USING CAA: AN EARLY EXPLORATION OF THE SLIM PILOT PROJECT

Helen Martin

Formative Assessment Using CAA: An Early Exploration of the SLIM Pilot Project

Helen Martin
School of Education
University of Aberdeen
Aberdeen
AB24 5UA
h.martin@abdn.ac.uk

Abstract

Supporting Learning in Mathematics (SLIM) pilot project: Can the research on formative assessment be applied to the use of a computer aided assessment (CAA) tool to enhance student learning?

The pilot project consisted of an initial online questionnaire and five 'study sessions' using QuestionMark Perception (QMP) with the BEd year 1 cohort at the University of Aberdeen. This short paper intends to stimulate a dialogue about how to meet the professional challenge of changing learner expectation and, in particular,

(a) How to design 'rich' questions

(b) How to provide meaningful feedback in a computer mediated environment.

by using a preliminary exploration of the data from the pilot project.

Introduction

There is an increasing concern about the level of what has been termed 'numeracy skills' in undergraduate students, not only within Initial Teacher Education (ITE) (Murphy (2005), Thwaites *et al* (2005)) but across academic disciplines in Higher Education Institutions (HEI) throughout the UK (Pidcock *et al* (2004), Agnew (2000)).

On a more global level, the increasing numbers and diversity of students entering Higher Education can lead to an acute pressure on resources (Sadler, 1997) and is forcing a critical and creative re-evaluation of how we respond to learner's needs. The associated increase in marking workloads and reduced contact time between staff and students can make it difficult to provide students with effective, regular and timely feedback on their performance. In this paper, the use of formative online assessment is explored as a means of enhancing students' learning by providing regular, detailed and constructive feedback on their learning.

This approach is strengthened by the recognition of 'formative assessment' as a means to raise standards within the school context (Black & Wiliam (1998), ARG (1999), Black *et al* (2002)) which has resulted in an unusual confluence of theory and practice in Scottish Education. It is further supported by evidence within Higher Education from authors such as Hounsell (2003) and Schmidt *et al* (1990). In order to try to clarify what is meant, in this paper, by 'formative assessment' a working definition is given below:

"Any process/activity which promotes learning by generating feedback information that is of benefit to students [and teachers] whilst engaged in the task itself; which enables the student to monitor continuously the quality of what is being produced and to develop their understanding / skills."

The emphasis is on ipsative-referenced assessments to encourage students to become more self-regulating (Yorke, 2003) rather than the preoccupation of marks / grades in order to compare or rank students where there is rarely an opportunity for the student to receive and act on feedback (Black & Wiliam (2003), Sadler (1989)).

The particular aspects of the research into formative assessment which seem to apply to the use of CAA are questioning techniques (QCA, 2003) and meaningful feedback (Sadler, 1998) since the other aspects are inextricably linked to synchronous dialogue. This dialogue between teacher and pupil is not easily transferred into an HE environment with the structure of large lectures, minimal contact time and reduced staff.

Wiliam (1999) refers to both 'rich questioning' and 'rich questions' although the former implies a dialogue between teacher and student where there are further exploratory questions, depending on the student's response, to elicit the underlying concepts. Within a computer mediated environment it is perhaps more appropriate to talk about 'rich questions' i.e. ones which 'illuminate aspects of student thinking rather than just measure attainment' (Black & Wiliam, 2003) where the responses and feedback provide the student (and teacher) with what they 'can do' as well as a diagnosis of errors in concepts and finding ways to address these. Watson & Mason (1998) are particularly interested in how to reframe questions to allow pupils to demonstrate higher order thinking skills and continue to develop a framework for effective questioning in mathematics in school classrooms specifically.

Feedback has the potential to improve learning and self-esteem however this is not always the case (Hyland, 2000). To be meaningful it should be more than a transmission of correct/incorrect with a worked solution provided. According to Sadler (1998) it should be specific to the task and the student's response to that task. It is not so much the quality of the feedback itself but rather the impact it has on the student; does it cause thinking?

The SLIM project outlined below is focused on formative assessment and, in particular, questioning and feedback.

Background

The pilot project (SLIM) commenced in Jan 2005 with BEd 1 students in mathematics. This collaborative project with the Learning Technology Unit (LTU), University of Aberdeen is intended to develop an online formative assessment tool, using QuestionMark Perception, which would allow students to develop their confidence and competence in Mathematics.

Our aims are to

- contribute to the development of a wider range of effective assessment in order to support our students;
- improve accessibility and feedback of formative assessment, especially in terms of the range of methods of assessment, in a manageable and practical way;
- support students to develop a level of independence and responsibility for their learning.

Outline

The pilot project consisted of an online questionnaire and five 'study sessions', each of which had 20 questions covering a variety of maths topics as well as some theoretical questions linked to the course inputs. The last question in each session was an opportunity for the students to provide feedback to the developers. The questioning techniques were influenced by the work of Wiliam, Watson and Mason as well as the particular functionality of the software used although the latter had a profound effect on the development of the questions.

Week 30	31st Jan	Initial Questionnaire
Week 31	7th Feb	Session 1
Week 32	14th Feb	Session 2
Week 33	21st Feb	Session 3
Students away for 7 weeks		
Week 41	18th Apr	Session 4
Week 42	25th Apr	Session 5

Involvement in the project was voluntary and anonymous;

Discussion

‘We cling to the familiar, like a much-loved old garment, even when, sometimes, it is long past its best and ought to have been discarded long ago’

(Broadfoot, 2001)

Perhaps what we should be working towards is ‘constructivist assessment’ (Roos & Hamilton (2005), Shepard (2001)) where the assessment is embedded within and an integral part of learning and teaching; where feedback is provided which supports the student’s own construction of an understanding.

‘If arguments in favour of formative assessment are to survive and prosper they must be articulated more fully and explicitly, and be built on more than taken-for-granted assumptions about what constitutes “good practice”’

(Torrance, 1993)

How can we meet the professional challenge of changing learner expectation?

Bibliography

AGNEW, C.T., (2000). Improving Numeracy Workshop. *Proceedings of the Annual Conference of the Royal Geographical Society with the Institute of British Geographers* 2000, Jan 5*.* University of Sussex, UK.

ASSESSMENT REFORM GROUP (ARG), (1999). *Assessment for Learning: Beyond the Black Box.* Cambridge: School of Education, University of Cambridge.

BLACK, P., HARRISON, C., LEE, C., MARSHALL, B., and WILIAM, D., (2002). *Working Inside the Black Box: Assessment for learning in the classroom.* London: Department of Education & Professional Studies, King's College.

BLACK, P. and WILIAM, D., (1998). *Inside the Black Box: Raising standards through classroom assessment.* London: Department of Education & Professional Studies, King's College.

BLACK, P. and WILIAM, D., (2003). 'In Praise of Educational Research': formative assessment. *British Educational Research Journal,* **29**(5), 623-637.

BROADFOOT, P., (2001). Editorial: new wine in old bottles? The challenge of change for educational assessment. *Assessment in Education, **8**(2), 109-112.*

HOUNSELL, D., (2003). No comment? Reshaping feedback to foster high quality learning. *Proceedings of Learning and Teaching Forum on Formative Assessment,* November 27. University of Edinburgh, UK.

HYLAND, P., (2000). Learning from feedback on assessment. In: A. Booth and P. Hyland, eds. *The Practice of University History Teaching.* Manchester: Manchester University Press.

MCALPINE, M., (2002). *Principles of Assessment.* Luton: CAA Centre.

MURPHY, C. (2005). The Role of Subject Knowledge in Primary Trainee Teachers' Approaches to Teaching in the Topic of Area. *Proceedings of the Sixth British Congress of Mathematics Education 2005,* March 30 - April 2. University of Warwick, UK.

PIDCOCK, D., PALIPANA, A. and GREEN, D., (2004). The role of CAA in Helping Engineering undergraduates to Learn Mathematics. Presented to *8th International CAA Conference 2004,* July 6 - 7. Loughborough, UK.
Available:
www.caaconference.com
[Date accessed: 21/12/04]

QUALIFICATIONS AND CURRICULUM AUTHORITY QCA, (2003). *Assessment for Learning. Using Assessment to Raise Achievement in Mathematics.* London: QCA Publications.

QUALITY ASSURANCE AGENCY for HIGHER EDUCATION (QAA), (2003). *Handbook for enhancement-led institutional review: Scotland*. Gloucester: QAA Publications.

ROOS, B. and HAMILTON, D., (2005). Formative Assessment: a cybernetic viewpoint. *Assessment in Education, **12**(1), 7-20.*

SADLER, D. R., (1998). Formative Assessment: revisiting the territory. *Assessment in Education, **5**(1), 77-84.*

SADLER, R., (1997). Assessment items: design, construct, use, refine. *Proceedings of computer assessment workshop,* February 14. Uniserve Science : University of Sydney.

SCHMIDT, N.G., NORMAN, G.R. and BOSHUZEN, H.P.A. (1990). A cognitive perspective on medical expertise: theory and implications. *Academic Medicine*, 65, 611-621.

SHEPARD, L. A., (2001). The role of classroom assessment in teaching and learning. In: V. Richardson, ed. *Handbook of Research on Teaching.* 4th ed. Washington: American Educational Research Association.

THWAITES, A., HUCKSTEP, P. AND ROWLAND, T. (2005). The Knowledge Quartet: Sonia's Reflections. *Proceedings of the Sixth British Congress of Mathematics Education 2005,* March 30 - April 2. University of Warwick, UK.

TORRANCE, H., (1993). Formative Assessment: some theoretical problems and empirical questions. *Cambridge Journal of Education*, **23**(3), 333-343.

WATSON, A. and MASON, J., (1998). *Questions and Prompts for Mathematical Thinking.* Derby: Association of Teachers of Mathematics (ATM).

WILIAM, D., (1999). Formative Assessment in Mathematics: Part 1: Rich Questioning. *Equals* **5**(2), 15-18.

WILIAM, D., LEE, C., HARRISON, C. and BLACK, P., (2004). Teachers developing assessment for learning: impact on student achievement. *Assessment in Education* **11**(1), 49 - 65.

YORKE, M., (2003). Formative assessment in higher education: Moves toward theory and the enhancement of pedagogic practice. *Higher Education,* **45**(4), 477-501.

ITEMBANKING INFRASTRUCTURE:

A PROPOSAL FOR A DECOUPLED ARCHITECTURE

Mhairi McAlpine and Linn van der Zanden

Itembanking Infrastructure: A Proposal for a Decoupled Architecture

Mhairi McAlpine and Linn van der Zanden
Scottish Qualifications Authority
Computer Assisted Assessment
The Optima Building
58 Robertson Street
Glasgow
G2 8DQ

Mhairi.McAlpine@sqa.org.uk
Linn.vanderZanden@sqa.org.uk

Abstract

The paper aims to provide a comprehensive outline of the elements which make up an Itembanking system and through the use of basic workflows and diagrams create a visual of the overall system and user interaction. In particular it will provide an overview of the proposed Itembanking Infrastructure that SQA is currently developing, and steps which have been taken towards its realisation. Our aims in developing this are to promote more flexibility in assessment, improve on access, increase efficiency, cost-effective processes, enhancement of validity and reliability and improve possibilities for feedback and reporting. The functionality of an Itembanking system will be explored in light of the ways that institutions may use such a technical structure along with the challenges and issues surrounding its implementation.

We have divided the system into four main elements:

- The itembank itself which stores the items and facilitates searching
- The item production elements which generate items suitable for entry into the bank
- The test delivery elements which control the delivery, marking and reporting of the results
- The test generation elements which control items being selected from the bank and concatenated into tests.

The paper will focus on a ‘reference’ diagram which will provide an overview of the elements and associated software, the relationships between them and the overall interaction of the system. Within the four main elements, sub elements will be identified; including the storage of items, the generation of items, item description, item delivery, marking, result processing, item analysis and test construction. These will be explored with a view to defining the functionality of each element independently to allow autonomous development – fitting in with a standards based decoupled system. Existing projects and recommended standards in these areas will also be highlighted.

Role profiles and workflows are discussed in terms of how different users may interact with the system and roles may be transferred onto an electronic banking system. Future plans to establish user requirements for each component of the Itembanking Infrastructure will be discussed in the conclusion.

Introduction

With recent developments in educational technology and emerging standards, item banking is coming to the fore as an efficient and cost effective method of recycling expensively produced examination material. Although awarding bodies have used itembanking for decades and computerised systems for over 30 years; the advent of XML and a standardised way of describing assessment data encoded in the IMS QTI specification - together with the technological possibilities opened up by large scale distributed systems, have given people confidence in the stability of this approach. SQA in particular has been itembanking in a paper form since the 1960s. In the 1980s this was enhanced by the introduction of a computer database to hold the item usage data. The migration of current paper-based itembanks within SQA to a computerised, internationally recognised format is now progressing and SQA is looking toward developing a technical and organisational infrastructure to support the surrounding activities such as test construction, item analysis, delivery and marking. The underlying aim is to achieve more flexibility in assessment; improved access; more efficient cost effective processes; enhanced reliability and validity; advanced possibilities for feedback and reporting; and the provision of both paper and computer based assessment.

Itembanking may provide an intermediate step between fully on-line assessment and fully paper driven assessment as the items from the bank could always be printed and distributed in a traditional format, before CAA was introduced. It is also noted that this may ease the introduction of CAA, providing a bedrock from which CAA may be launched, with the major burden of this intermediate change being borne by the SQA rather than on candidates and centre staff. This may also prove a beneficial process for HE institutions. Although computerised itembanking is nothing new, existing systems tend to be monolithic entities with fixed functionality. Revamping the system features or adding additional is prohibitively expensive, however as we enter the brave new world of computerised assessment – ensuring that we take advantage of the increases in assessment approaches and validity enhancements that this will bring requires a flexible technological architecture.

This paper proposes a decoupled architecture, based on international standards and a webservices approach to the integration of functionality. The paper envisages the itembank itself as a unit made up of two components; a database which facilitates metadata storage, retrieval and search functionality, and a repository which facilitates the storage of items, resource files and manifest files. However, this bank is merely the datastore for a larger system which sits around the bank, feeds into it, interrogates it and exports from it.

Rationale for a Decoupled System

The concept of a decoupled, open source and standards based infrastructure is becoming increasingly popular[1]. For example, the JISC e-Learning Framework (http://www.elframework.org) adopts such a model in providing a networked service, typically using Web Services[2], referencing open specifications and standards that can be used to implement the service, and providing open-source implementation toolkits.

Diagram 1 suggests a potential architecture for a decoupled itembanking system. At its centre is the core itembank – comprised of a linked database and repository together with content unpackaging functionality. To the right are services associated with the generation and input of items; to the left are services associated with the export and delivery of items and at the bottom area are services associated with test construction.

Most existing systems conflate a number of these pieces of functionality into one software system, locking a user in to one provider and therefore requiring compromise to achieve the best overall fit. Decoupling the system in this way facilitates a "mix and match" approach. So long as each element inputs and outputs data in accordance with international standards and specifications, the integration of vendors, open source and custom built solutions are possible. In this system, elements may be replaced on a rolling basis with continual evaluation and thus providing the capability to keep up with new assessment approaches without periodic major upgrade.

[1] IBM (Leymann, Roller, and Schmidt, 2002) uses the service-oriented architecture (SOA) approach, which is the latest in a long series of attempts in software engineering that try to foster the reuse of software components and where programs are broken down into smaller programs through functional decomposition.

[2] An official standard for WSDL (Web Services Description Language) was released in 2001 by the World Wide Web Consortium. For a further description of web services see http://www.ariadne.ac.uk/issue29/gardner/.

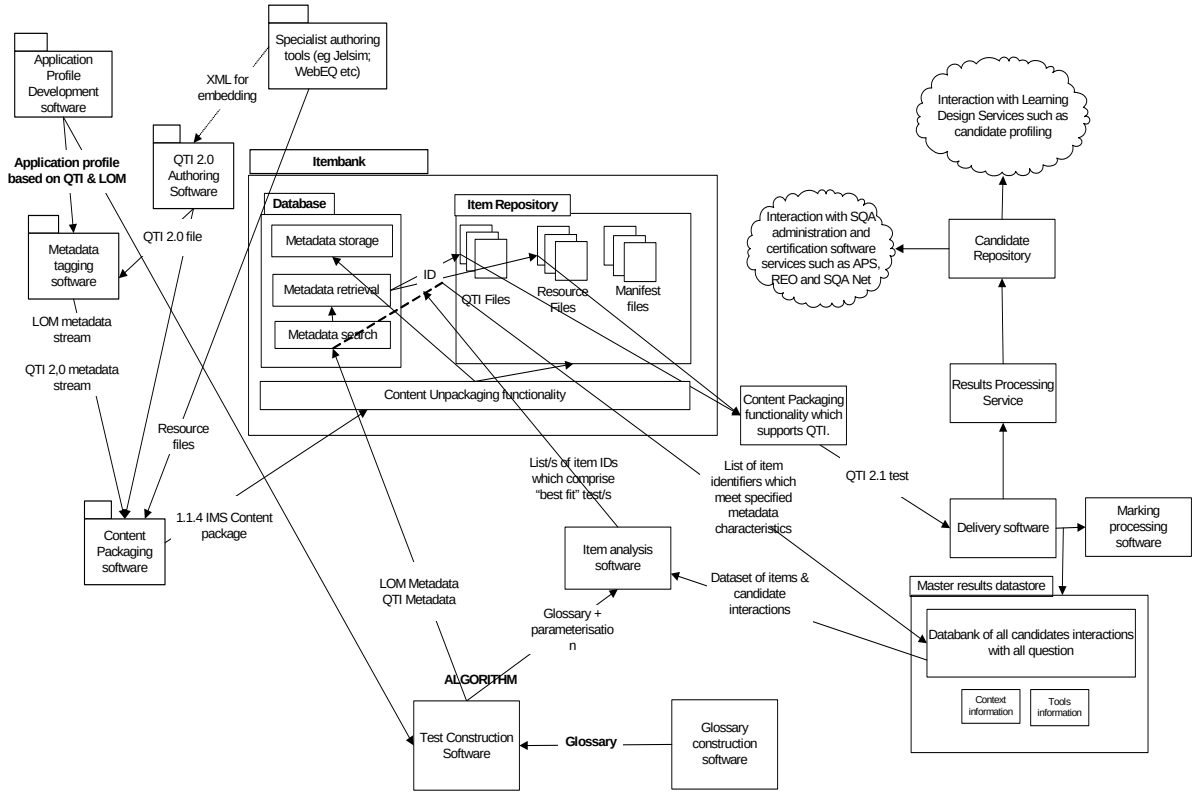


Diagram 1

Particular advantages of a decoupled architecture include:

- It can be easily adapted to accommodate a change in the model or workflow processes.
- A central team within an institution should be able to control the changes in workflow and software packages without major impact on the user.
- With a modular approach small chunks can be built and used immediately, with existing processes used to fill in the gaps until the next pieces are built.
- As the specifications grow and develop, pieces can be upgraded in line. Furthermore it can be developed cross-institutionally ensuring community involvement.
- As CAA 'beds-in' and people become more sophisticated in the way they use itembanking, additional demands will be placed on the system. A modular architecture allows for these demands to be slotted in at the appropriate points.

However the consequences and implications of adopting a modular approach include:

- The need to ensure standards compliance. Not just to the strict specifications, but where the specifications are loose to ensure that the manner implemented is in line with existing, emerging best practise.

- As a workable system will not be developed in one go, manual processes and pre-existing software systems will have to be built into the workflow. This may require developing additional functionality that will not be required once the full system is in place.
- Where a piece of the architecture is faulty, the possibilities for computer interaction with little user input may lead to difficulties with early detection of errors. System testing of each piece must be highly robust before it goes live.

What does Itembanking Entail?

We have divided the system into four main elements:

- The itembank itself: which stores the items in the repository and facilitates searching through the linked database
- The item production elements: which generate items suitable for entry into the bank
- The test delivery elements: which control the delivery, marking and reporting of the results
- The test generation elements: which control items being selected from the bank and concatenated into tests

The following section will identify the associated components to each element, a brief overview of the functionality all components, together with references to existing projects in this area and standards used.

Storage of Items

Within the proposed decoupled architecture the storage of the items is facilitated by two elements: the database and repository. The repository stores the QTI files, any associated resources from those files and the manifest files from the imported content packages as well incorporating content 'un-packaging' functionality in order to de-aggregate the elements of a submitted content package and deposit them into the repository and database as appropriate. The database stores both the metadata and QTI metadata, as well as housing the search and retrieval functionality.

The separation of the repository from the database allows for faster processing of search and retrieval requests. This is because the search is carried out on the metadata first before the item files are retrieved from the repository on the basis of the ID's of the selected items. Another advantage is the afforded greater control of security and access permissions, as the metadata can be made more available than the actual item files in the repository.

Examples of existing systems incorporating itembank activities are the Hamlet itembank used by BTL among others; the TOIA system developed at the University of Strathclyde and the itembanking functionality incorporated into Perception from Questionmark. These are, however, integrated closed systems.

Generation of Items

The software associated with the generation of items for an itembank would depend on the content of these items. The IMS Question and Test Interoperability (QTI) specification is a standardised format for exchange of assessment item data –making it the most desirable form for storage and hence native authoring output. There may also be a requirement for additional specialist software associated with authoring items which have particular requirements – such as the inclusion of mathematical notation or multimedia elements.

Once the questions themselves have been authored, they must be tagged with standardised metadata to facilitate the search and retrieval processes. The international standard for describing learning objects is the IEEE LOM, however the full specification is very extensive. Application profiles can simplify data entry particularly where there are a large number of similar questions being entered at once, increasing both speed and accuracy of metadata entry. Such profiles could be generated by an additional piece of architecture.

The QTI Authoring Software allows questions to be created and exported in QTI 2.0 format and is complimented by Specialist Authoring Software. This develops parts of the items which cannot be directly encoded in QTI 2.0 but are instead either embedded or called from within the item. The Metadata Tagger enables metadata to be entered and attributes data to the items held in the bank, such as the author, the subject area of the question and the type of item. Application Profile Development Software facilitates the development of application profiles (or customised templates) based on the LOM, which pre-fill or restrict the entries that are allowed into the fields.[3] The Content Packager packages together the elements of QTI 2.0 items according to the specifications of the IMS content packaging guidelines, facilitating import into any repository which recognises such standards.

Until recently no software capable of supporting the above activities has been available. The JISC-funded SPAID project (Young, MacNeill, Adams, McAlpine, 2005) produced a number of these as part of the Toolkit strand. The SPAID Metadata Tagger facilitates the generation and tagging of both LOM and QTI 2.0 Metadata. This application is customisable through the use of application profiles, while the SPAID Content Packager content packages assessment items in accordance with the QTI 2.0 specification. These are however very much at a prototype stage and further work is required to make them operational.

Delivery, Marking and Result Processing

Delivery, marking and result processing are post-itembank activities. Once tests are constructed they are passed into the delivery system, which then presents these assessments to the candidate and passes the input to the Marking Processing Software. The marked items are fed through the Result Processing Service which informs the Candidate repository and may interact with additional Services such as candidate profiling or administration and certification software. The Delivery system

[3] although it is recommended that the entire data schema is implemented even if many elements remain hidden to the end user, in order to ensure interoperability.

additionally submits all candidate interactions with items to the Master Results Databank for archiving.

The Delivery Software imports the assessments from the itembank in the form of a QTIv2.1 package. On completion of the assessment, the delivery software sends the recorded responses to the Marking Processing software. This consists of several elements which each facilitate the processing of different item types. There are three major approaches to mark processing; the first marks items entirely automatically, the second refers the items to a system where they are entirely human marked and the third uses a mixture of computer based and human marking. Question types which are best marked entirely by computer include Multiple Choice, Multiple Response and hotspot questions- each with their individual response processing template. Questions to be human marked, such as essays, would include a human readable mark scheme, while those using a mixed model- either human marked with a computer check or computer marked with human support- would use both.

The Result Processing Service software aggregates the marked items according to the requirements of the qualification, implementing the pass mark or grade boundaries which may be in force, while the Master Results Databank holds all the candidate interactions with items which are fed out from the delivery software – interacting with the item pools selected from the algorithms produced below.

New forms of marking are anticipated as CAA becomes more sophisticated and the flexibility to change or expand on response processing templates is desirable. The extracting and holding of candidate interactions would allow for more sophisticated analysis and process data, beyond processing scores of candidates on the items.

Test Construction

Test construction is the method by which items are concatenated to produce a test conforming to a particular specification. This has two aspects, firstly metadata searching to identify the questions which meet the descriptive metadata, then statistical analysis of the items and selecting those which meet specified parameters. A list of items comprising a suitable test are then sent back to the bank and a test file is exported for consumption by the delivery software.

Glossary Development Software would produce a glossary which defines the statistics to be used in the test construction system, providing the basis for the item analysis to take place, outputting a glossary in a standardised QTI format**.** Test Construction Software consumes application profiles together with the glossary to produce an algorithm, comprised of metadata (both LOM and QTI) and statistical terms which define the rules for test construction. These are then split – with the metadata first being sent to the bank, identifying an item pool that meets the defined criteria. The items are then matched with candidate interactions from the Master Results Databank, to produce a dataset which is sent to the item analysis software together with the statistical conditions from the algorithms. The Item Analysis Software runs the required analyses from the algorithm, identifying items from the pool that meet the conditions. Those items which meet the conditions of the algorithm are then passed back to the itembank for retrieval and packaging into tests.

The provision of the proposed architecture would increase reliability of item analysis, efficiency, and cost savings through reuse of items. The notion of the service

orientated infrastructure includes the generation of usage data at run time. It is envisaged that no pre-testing would be necessary, but instead a small number of live items would be constantly tested. This would negate the requirement to hold static statistics for test generation purposes.

Although there is currently no existing Itembanking software which includes sophisticated test construction and usage data capture, there are a number of Item Analysis software packages available. The CATS project (Tulloch, 2006) is creating a toolkit to support automated assessment construction. It will build upon the outputs of two previous ELF projects – SPAID (Storage and Packaging of Assessment Item Data) and Discovery Plus (D+ - Brokerage for Deep and Distributed e-Learning Resources Discovery). The overall aim of the project is to create a toolkit of loosely-coupled web services which support the various tasks inherent to automated assessment construction e.g. searching for, retrieving and aggregating assessment items held in multiple item banks.

Roles and Workflows

This area is yet to be defined, in particular with regards to an overall system. Previous attempts to capture user roles and processes include the IBIS report (Sclater, 2004) and User requirements for the ultimate online system (Sclater and Howie, 2003). Each of the four elements discussed throughout the paper would have different users interacting with them.

When considering user roles, it should be noted that individuals may play one or more roles, both in the existing paper based system and in an online system. To facilitate adequate allocation of permissions however, it is necessary to exhaustively define roles – linking them with bank access and actions. These actions need to be clearly and exhaustively defined before linking with roles.

Established roles within our existing infrastructure include item writers (who write assessment items), qualifications staff (who oversee the assessment administration) and principal assessors (who oversee the assessment process), within those major roles however, there are a number of different functions that they and others perform. These roles may change or be separated into different functions within an operational itembanking system.

Workflow processing also requires further consideration, although some workflows on the generation of items have already been suggested in systems. One of the advantages of a decoupled architecture however, is that workflow processes may be changed as demands placed on the system change over time.

Where workflows are transferred onto an itembanking system, some elements of the workflow will be eliminated, other elements modified and new elements introduced. Workflows for producing a test from an itembanking system may include authoring and moderation of items, authoring and validation of metadata, and manual moderation of an automatically constructed test.

Business Process Execution Language (BPEL) may be used to orchestrate and manage the workflows in which partners in the process are identified and declared, the workflow is designed and defined, and business logic is added using BPEL Constructs before validation and deployment take place. This will produce a clear overview and relationship between the processes in each element of the Infrastructure.

Populating and Monitoring the Bank

Once the infrastructure is developed, the bank will need populating before operational use; consideration of the issues around population and monitoring is desirable at this stage to highlight issues which may impact on the software design.

Question writing procedures will have to be reconsidered to support a banked system. The most popular method of bank population involves content grids. However an extensive system as is being planned here may need a more sophisticated approach. Consideration should be given to parameterised questions, to reduce the impact of any potential security breaches and item exposure effects. This has implications for item analysis and may prove too sophisticated to handle in the short term.

A review is required to determine acceptable item exposure rates and extrapolate minimum bank size for each content area identified. This will help to inform the power of the search functions required and the space needed for storage.

One of the major advantages of banking items is the increased quality assurance it affords. Mechanisms for monitoring and regulating item exposure as well as procedures for discarding/quarantining over exposed items should be developed. Curricular drift (where a content area goes in or out of educational fashion) should be monitored and addressed through the dynamic item analysis including the use of item trend lines. Any unjustified deviation should be monitored and flagged to the relevant subject teams to ensure construct validity.

Problems may arise in banking where items are not independent for example where they have the same source paragraph, or refer to one another. Although complex banking can overcome these issues they require careful consideration. As a preliminary stage, these dependencies should be eliminated as much as possible, or the group should be banked as a single item.

Conclusion

This is an outline of a proposed system to facilitate sophisticated electronic itembanking using a webservices model to enable a decoupled system. Beyond the overview of services and their interactions, the precise definitions, requirements, interactions and data transfers for each of the elements must be scoped. The roles of users interacting with the system need to be defined and the workflow processes likely to be used must be identified.

An overview of the elements is given in this paper and can also be found in McAlpine et. al. (2006), while further work on defining the precise requirements for each of the elements as well as work on user roles and workflows is ongoing within SQA at the moment.

The Scottish Qualifications Authority appreciates the need to engage with the educational community from nursery to higher education to ensure that our technological structures are supportive of the forms of assessment that can best support learning and teaching. We see sophisticated itembanking structures as one component of a well-rounded modern assessment system and are keen to engage with all sectors in scoping, developing and evaluating a system which will be of benefit to all.

References

F. *Leymann*, D. *Roller*, and M.-T. *Schmidt*. Web services and business process management, IBM Systems Journal, Volume 41, Number 2, 2002.

Available at *http://www.research.ibm.com/journal/sj/412/leymann.html*

I. Tulloch, J. Everett, R. Young, M. Watson and R. Taylor. CATS- Constructing Assessments using Tools and Services, 2006.

N. Sclater. IBIS Item Banks Infrastructure Study. HEFCE, 2004.

N. Sclater and K. Howie. User requirements of the ultimate online system Computers & Education 40, 2003

R. Young, S. MacNeill, D. Adams, M. McAlpine. SPAID Final Report, 2005. Available at http://www.jisc.ac.uk/uploaded_documents/SPAIDfinalreport.doc

LIGHT-WEIGHT CLUSTERING TECHNIQUES FOR SHORT TEXT ANSWERS IN HUMAN COMPUTER COLLABORATIVE (HCC) CAA

**Mary McGee Wood, Craig Jones,
John Sargeant and Phil Reed**

Light-weight Clustering Techniques for Short Text Answers in Human Computer Collaborative (HCC) CAA

Mary McGee Wood, Craig Jones, John Sargeant and Phil Reed
School of Computer Science,
University of Manchester
mary@cs.man.ac.uk
jonesc@cs.man.ac.uk
js@cs.man.ac.uk
preed@cs.man.ac.uk

Abstract

We first explore the paedogogic value, in assessment, of questions which elicit short text answers (as opposed to either multiple choice questions or essays). Related work attempts to develop deeper processing for fully automatic marking. In contrast, we show that light-weight, robust, generic Language Engineering techniques for text clustering in a human-computer collaborative CAA system can contribute significantly to the speed, accuracy, and consistency of human marking. Examples from real summative assessments demonstrate the potential, and the inherent limitations, of this approach. Its value as a framework for formative feedback is also discussed.

Introduction

Assess By Computer (ABC; Sargeant et al 2004), deployed at the University of Manchester since 2003, follows a human-computer collaborative (HCC) approach to assessment. We focus on constructed answers such as text and diagrams rather than answers requiring mere selection between alternatives. The HCC assessment process is an active collaboration between humans and a software system, where the software does the routine work and supports the humans in making the important judgements.

One feature which distinguishes our approach from “traditional” CAA is our classification of question and answer types, which has three parameters. First, we distinguish constructed from selected answers (we strongly deprecate the traditional use of the term “objective” to mean “selected”).

Second, we distinguish “closed” or truly “objective” from “open” or “subjective” questions. For closed questions, the substance of a correct answer can be specified in advance (although its expression can vary wildly and unpredictably: Wood et al 2005). Open questions typically ask for an original example or argument. A marking scheme can only describe meta-level properties of a correct answer, and a “model answer” can only be an example.

Third, we distinguish loosely between long and short text answers. Length does not necessarily correlate with openness /closure: “Describe the causes of haemolytic disease in the newborn” calls for a paragraph of routine book-work while “Give an original example of an exception to default inheritance” requires only a short phrase. Length also does not necessarily correlate with the levels of Bloom’s taxonomy (Bloom et al 1956). Its main significance in ABC is that different Natural Language Engineering techniques are optimised for different lengths of text. To date we have focussed on simple, robust, generic techniques which are best suited to short answers.

Related Work

The use of text clustering in CAA is far from unique; but the other work we are aware of, such as the examples below, limits itself to formative assessment and/or aspires to be fully automatic.

Lütticke (2005) uses “logical inference” to compare student-drawn semantic networks with a model answer and generate formative feedback: the details of the comparison mechanism are unclear.

Weimer-Hastings et al (2005) use Latent Semantic Analysis to compare student answers with expected answers in an Intelligent Tutoring System in research methods in Psychology. Its use is purely formative, and they have attempted to evaluate student learning gain but not the effectiveness of clustering per se (p.c.). Although the technique is generic, its application is question-specific: they refer to it as “expectation-driven processing”.

Carlson & Tanimoto (2005) induce text classification rules from student answer sets. These rules are used “to construct ‘diagnoses’ of misconceptions that teachers can inspect in order to monitor the progress of their students” and to automatically construct formative feedback.

Pulman & Sukkarieh (2005) aim for automatic marking of “short” (“from a few words up to five lines”) free text answers to factual (objective, in our terminology) science questions. They use relatively heavy-weight techniques from traditional computational linguistics, and compare answers with keyword-based “patterns”, for which machine learning techniques have been investigated. They have worked with real student data, and their best results correlate acceptably with human markers’ judgements, but on a very small sample, and it is not obvious that these techniques will scale up sensibly.

The Paedogogic Potential of Short Text Answers

Constructed-answer questions have significant advantages over selected-answer questions for assessing students, even at the “knowledge” and “comprehension” levels of Bloom’s taxonomy. Recalling even a bare phrase like “mean cell volume” is a greater challenge than recognising it, even among cunningly chosen distractors; let alone the possibility of getting it right by luck.

And even short text answers (1-30 words; or comparably simple diagrams) are surprisingly versatile. As the following examples (with genuine, representative, mostly good student answers) show, short text answer questions, set cleverly, can test all levels of the taxonomy.

Knowledge: *What single measurement would you make to confirm that an individual is anaemic?*

Student answer: *haemoglobin concentration*

Comprehension: *A blood sample was taken from a patient and he was found to have a high white cell count. On further investigation the patient was found to have a neutrophil count of 22 x 10^9/L. Give two examples of what this could be indicative of.*

Student answer: *A recent or present bacterial infection. Or an allergic reaction.*

Application: *What is the value at the root of this minimax tree?*

Student answer: *42*

Analysis: … *What general significant problem with the size of search spaces does this illustrate?*

Student answer*: There are too many to calculate. This problem illustrates the number of possible choices AI problems have to deal with; it is a combinatorial explosion.*

Synthesis: *Rewrite the following replacing the underlined part with the appropriate pronoun: Ho regalato I quaderni a Paolo.*

Student answer: *Glieli ho regalati*.

Evaluation: *For each of the following pairs of classes, state whether or not it would be appropriate to relate them by inheritance, and why. If not, what other sort of relationship would be appropriate? – Car and Wheel*

Student answer: *This one may be better as a composition instead. A car as an association with wheel, but a wheel can exist on its own without the car class.*

Text Clustering

Clustering is the process of grouping similar objects together. A measurement of similarity, or distance, is used to assign objects within a set into subsets or clusters. Clustering is used in other fields such as Bioinformatics (Heyer et al 1999), finding nearest neighbours of a document (Buckley & Lewitt 1985), and for the organisation of search engine results (Zamir et al 1997).

Clustering offers a number of benefits in HCC assessment. The examples used here are free text student responses to assessment questions. Similar work at Manchester using the ABC system is looking at diagram responses (Tselonis et al 2005). Clustering similar answers together can help the human marker, as it provides a review mechanism to check that marking is consistent, and potentially offers a basis for rapid formative feedback.

The simplest form of text clustering is based on keywords, which may be specified in advance or (according to the HCC approach) expanded during the marking process. This has proved useful in some cases (as shown below), but is not a general solution. In this paper we concentrate mainly on the consequences of clustering the complete texts of short answers.

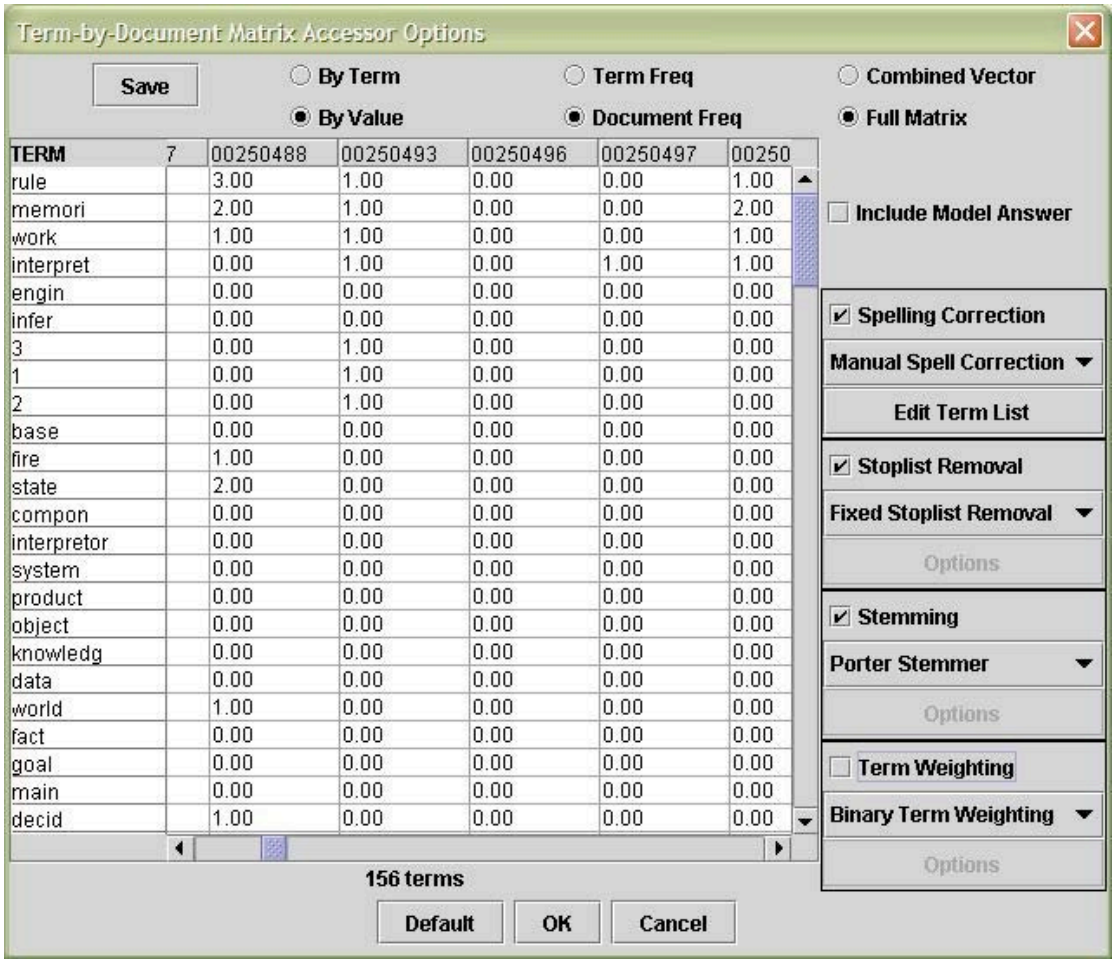
Clustering offers a tradeoff: the larger the clusters, the more fewer there are to process, but the less similarity there is between answers within a cluster. For formative applications we may be able to live with some inaccuracy in order to be able to give rapid feedback per cluster. In the summative case very high standards of accuracy are required if the students are to have confidence in the assessment software and procedures.

Lightweight Clustering Techniques

A commonly used measure of similarity from the field of Information Retrieval is the Vector Space Model (Salton 1971). Documents are expressed as vectors within a multi-dimensional space. The similarity between two documents is calculated as the distance between their respective vectors.

This clustering process can be broken down into a number of distinct steps, which have been implemented within a prototype extension of the ABC marking tool. The first step is the creation of a *term-by-document matrix*, a list of terms (words) and a count of the number of times they appear in each answer (see Figure 1). Each column is a vector representing the term frequency counts of an individual answer. Several pre-processing steps can be performed on the matrix to improve performance. These include spelling correction, removal of stop words (commonly occurring words of little interest such as *"the"*), stemming (removal of affixes from a word to leave a common stem. e.g. *"interpreter"* is converted to *"interpret"* - Porter 1980) and applying different weights to terms, in our case binary.

The next step is to calculate the similarities between vectors. The simplest way is to take the Euclidean distance between vectors. However this does not normalise vectors for length, and so the measure commonly used is the cosine of the angle between two vectors. This gives a range between 0.0 and 1.0, where a value of 0.0 indicates two answers that share nothing in common, and a value of 1.0 indicates two answers that are identical after pre-processing. This similarity measure can then be used to cluster the answers.



TERM	7	00250488	00250493	00250496	00250497	00250
rule		3.00	1.00	0.00	0.00	1.00
memori		2.00	1.00	0.00	0.00	2.00
work		1.00	1.00	0.00	0.00	1.00
interpret		0.00	1.00	0.00	1.00	1.00
engin		0.00	0.00	0.00	0.00	0.00
infer		0.00	0.00	0.00	0.00	0.00
3		0.00	1.00	0.00	0.00	0.00
1		0.00	1.00	0.00	0.00	0.00
2		0.00	1.00	0.00	0.00	0.00
base		0.00	0.00	0.00	0.00	0.00
fire		1.00	0.00	0.00	0.00	0.00
state		2.00	0.00	0.00	0.00	0.00
compon		0.00	0.00	0.00	0.00	0.00
interpretor		0.00	0.00	0.00	0.00	0.00
system		0.00	0.00	0.00	0.00	0.00
product		0.00	0.00	0.00	0.00	0.00
object		0.00	0.00	0.00	0.00	0.00
knowledg		0.00	0.00	0.00	0.00	0.00
data		0.00	0.00	0.00	0.00	0.00
world		1.00	0.00	0.00	0.00	0.00
fact		0.00	0.00	0.00	0.00	0.00
goal		0.00	0.00	0.00	0.00	0.00
main		0.00	0.00	0.00	0.00	0.00
decid		1.00	0.00	0.00	0.00	0.00

Figure 1: A Term by Document Matrix

Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering (Jain et al 1999) starts with each object forming a separate cluster. The process then follows these steps.

1. Find the two most similar clusters, A and B.
2. Combine A and B into one cluster.
3. Repeat until a designated stop point.

The ultimate end point is a single cluster that contains all answers. This is uninformative. One of our most interesting questions is how to determine the most effective stop point for a given question for a given purpose, bearing in mind the speed / accuracy trade-off discussed above.

In the initial state it is straightforward to calculate similarity between clusters, as they each contain only one item. Similarity between clusters containing multiple answers is more complex. “Average linkage” (the mean of the distance between all elements within cluster A and cluster B) is commonly used as a measure.

Within Cluster Similarity is a measure of how similar answers are to each other within any given cluster. Average Within Cluster Similarity is a measure

of how good the clusters are overall. A value approaching 1.0 indicates answers within each cluster are highly similar to each other.

Experimental Design

The data used here comes mostly from first year undergraduate summative examinations in Artificial Intelligence in the School of Computer Science (although ABC assessments have been run in a variety of subject areas, including Italian, Linguistics, and Pharmacy). All answers shown here have been marked by a human assessor.

Similarity between answers was calculated using the Vector Space Model as outlined above. The clustering algorithm was run to each of three termination points, which we believe (on the basis of experience) can produce useful clusters. *Optimal* termination points will vary among questions and assessment modalities, further reinforcing the tenet of HCC that some control must reside with the human marker.

The first termination point is to take the last clustering step when the Average Within Cluster Similarity value is equal to 1.0, indicating that all answers within each cluster are identical after pre-processing. The second is to cluster to a value of Average Within Cluster Similarity of 0.95. At this point answers within a cluster are not functionally identical to each other, but should still be reasonably similar. The third is to examine clustering from an efficiency aspect, considering how much effort could be saved for the human marker if marking by cluster were to be safe. For this we took a point when the number of clusters is 50% of the initial number of answers.

Examples

Experience in marking reveals three categories of question and answer: those where we can mark fairly consistently by cluster, those where marking by cluster is unsafe but reviewing marks by cluster is valuable, and those where clustering buys us little or nothing.

Answers where we can consistently mark by cluster

This knowledge-level question responds well to clustering:

CS141204 q1.1a*. In the "Hector's World" lab, conflict resolution is handled by "salience". Name two other conflict resolution strategies which can be used in production systems*.

Model answer: *Any two of rule ordering, specificity, recency, random.*
NB priority is not acceptable, as it is a synonym for salience.

Partial analysis at the limit of Average Within Cluster Similarity 1.0:

Cluster 1 (“Specificity”, “Random”): 13 answers, Mark = 2
Cluster 2 (“Specificity”, “Rule Ordering”): 8 answers, Mark = 2
Cluster 3 (“Specificity”, “Source File Ordering”): 6 Answers
    Mark = 2: 5 answers
    Mark = 1: 1 answer (“Specification” – error of stemming)
Cluster 4 (“Source File Ordering”, “Random”): 5 answers, Mark = 2
Cluster 5 (“Specificity”, “Priority”): 5 answers
    Mark = 2: 2 answers
    Mark = 1: 3 answers
…
Outliers = 67 answers

The anomaly in Cluster 5 is due to human error by the marker. The version of the ABC marking tool used for this exam did not yet incorporate clustering: had it done so, this mistake would have been avoided. As shown in column 4 of Table 1 in the Appendix, when clustering is continued to the point where the number of clusters is half the number of answers, 4% of the answers have marks different from the rest of their cluster – an acceptable level of accuracy for some types of assessment, and certainly for formative feedback by cluster..

Further clustering improves efficiency, but at a corresponding cost to accuracy. Answers missing correct terms, or with incorrect terms, are merged with correct answers if the clustering process is taken too far.

The fact that clustering collapses word-order can provide useful generalisations, as it does for this question. For another knowledge level question,

CS141205 q3.1a: *The CS1412 "Hector's World" lab uses the programming environment JESS. What does "JESS" stand for?*

Model answer: *Java Expert System Shell*

some students answered “Java Expert Shell System”. These were clustered with “Java Expert System Shell”, and were marked as correct by the human.[1] And BL181104A Q 1.1 ``What single measurement would you make to confirm that an individual is anaemic?" returned, as its fourth largest cluster, 13 minor variants on “haemoglobin concentration in the blood”, comprising 11 distinct text strings which had been correctly collapsed by pre-processing (see Figure 2). (We will see below, however, that there are other questions for which word order information about the answers is needed.)

[1] (As they were, compared to such outliers as “Java Encapsulated System Software”, “Java emulator simulator system”, or “Java Expressions Structurated System”.

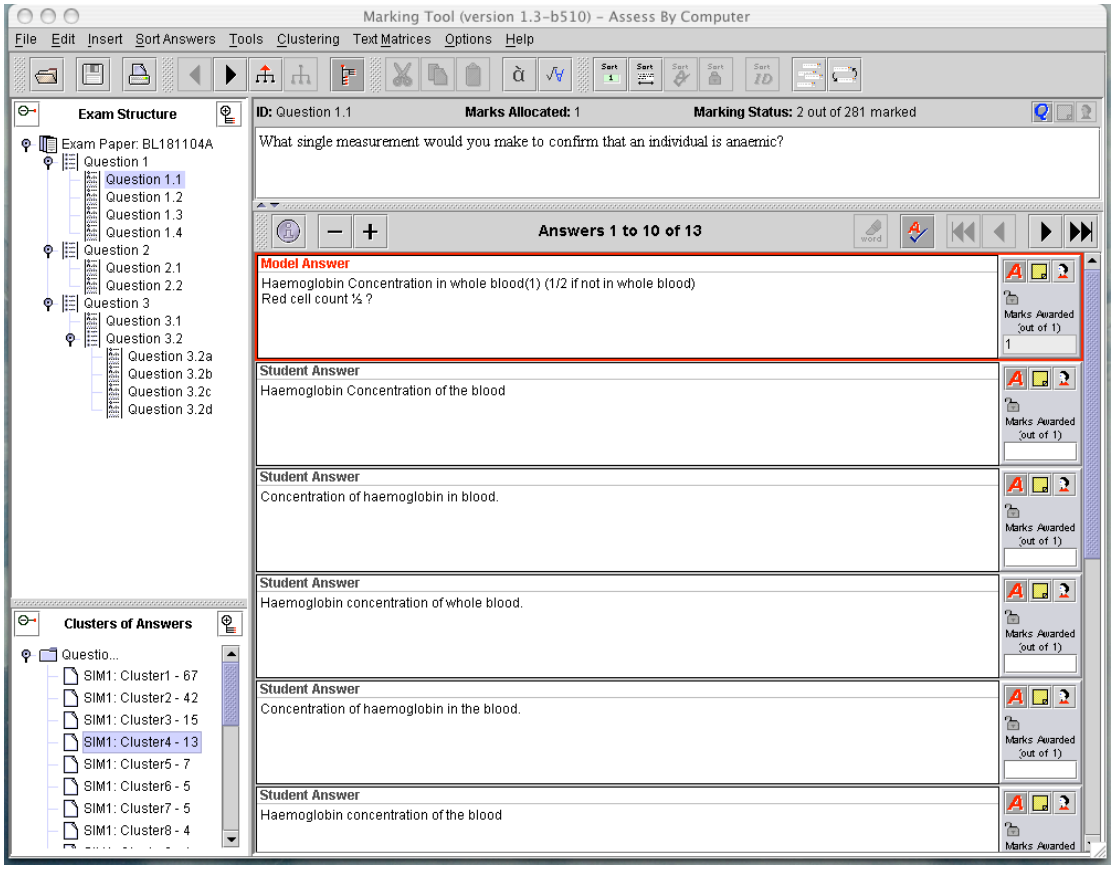


Figure 2: *anaemia*, Agglomerative Hierarchical clustering

Clustering used as a review

The following example shows a type of question which is more difficult to mark:

CS141205 q1.2: *A cricket ball used in a day-night match is white. What problem does this cause for semantic networks? Give another example of the same problem (but **not** the example featured in the lectures).*

Model answer: *Exceptions to default inheritance. Anything sensible except penguins as non-flying birds, since that was the primary lecture example*.

Answers with a high degree of similarity for the first part of the question might have different responses to the second; or vice versa. The question also asks for an original example. As a result the answers are highly variable. So are the marks awarded to answers within a cluster (see column 6 of Table 1: 22% of answers have marks anomalous for their cluster). Clustering is most useful for the second part, identifying similar examples (especially those students who ignored the question and used penguins as an example).

At Average Within Cluster Similarity 0.95 (91 answers, 73 clusters):

Cluster 1, 4 answers: (non-flying birds, 3 penguins, 1 ostrich)
Cluster 2, 3 answers: (3 wheeled cars)
Cluster 3, 3 answers: (non-flying birds, 2 penguin, 1 chicken)
…
Outliers, 60 answers (also includes 3 wheeled cars, ostriches)

Here clustering by keyword comes into its own (see Figure 3). Answers using the word "multiple" demonstrated a predicted common misunderstanding and were awarded, at most, one mark for a good example.[2] Any answer containing "exception" was awarded full marks unless it also contained the word "penguin".

Keyword="multiple": 13 answers
Mark=0: 9 answers (misunderstanding)
Mark=1: 3 answers (good example, first part wrong)
Mark=2: 1 answer (see footnote)
Keyword="exception" & NOT "penguin": 30 answers
Mark=1: 3 answers (bad examples, 1 chicken)
Mark=2: 27 answers

While marking by cluster is dangerous for this type of question, clustering is still of some benefit in allowing a human user to review their marking judgments, and may offer a useful basis for per-cluster formative feedback.

[2] With one exception: "A traditional cricket ball is red. If the semantic network has defined a cricket ball it as red, then multiple hierarchy will be needed for a white the ball to define a different type of ball with colour white.

Another example of this would be Manchester United players IsA Footballer, Manchester United players have Skill: High, Intelligence: High. Phil Neville IsA Manchester United player. Skill: Low, Intelligence: Low. It does not fit the normal semantic network for a Manchester United player."

The first part of this answer is wrong; but the human marker (a Stockport County supporter) awarded a bonus mark for the originality of the second.

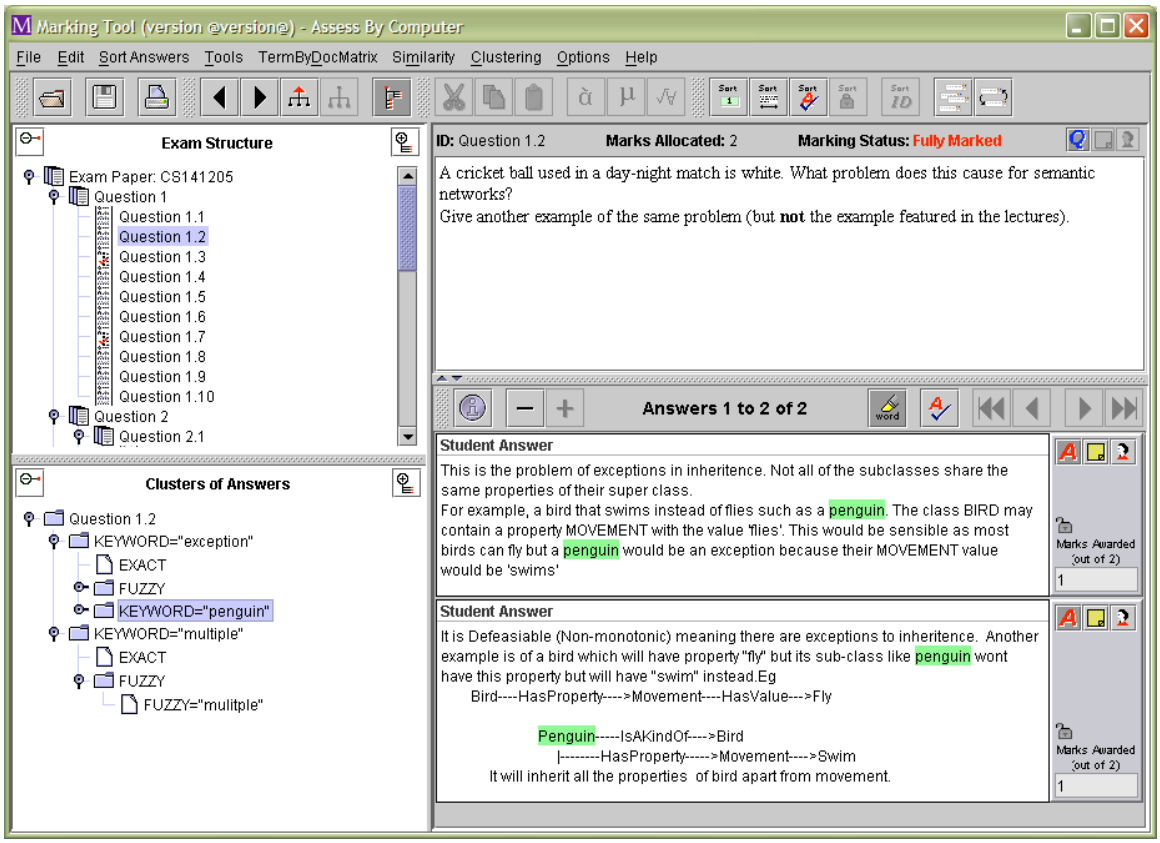


Figure 3: *penguins*, clustering by keyword

Where clustering can't take us

Thus far we have considered question types where correctness was determined by the content of the answer. Any vector-based approach must fail where correctness is a meta-level property of the *structure* of an answer. Consider this example:

CS141205 Q1.1. *A traditional cricket ball is red. Express this fact as a very simple semantic network, in two different ways.*

Model answer: *cricket ball --<has property>-- colour --<value>-- red*
cricket ball --<has colour>-- red

Clustering 93 answers into 63 clusters (with average within-cluster similarity 0.97), we find this, clustered with four correct answers:

Cricket Ball → HAS-COLOUR -→ Red
Cricket Ball:
Colour Red

The clustering is based on the word "has-colour". This answer is wrong because the two "networks" are not sufficiently different from each other (as can be seen by comparison with the model answer). It is inherently impossible

for any clustering technique based purely on word occurrences to detect this. More sophisticated techniques would be more expensive and more fragile.

As with the penguins, the keyword manager can be useful here – all the answers including (variants on) “has-value” received full marks. This reinforces our position that light-weight techniques manipulated by human intelligence offer a viable and valuable strategy for CAA.

A less significant weakness of vector-based clustering approaches is that word order is not taken into account. In some cases this is acceptable or even advantageous, as shown above. However consider the following:

CS141203 q1.1a: *… What conflict resolution strategy would you use to force rule 2 to fire? What strategy would you use to force rule 3 to fire?*

Model answer: *Rule 2 – specificity. Rule 3 - priority*

The answers are short, but clustering shows a much lower correlation with human judgement than for the previously analysed questions. This is largely because the incorrect answers “priority, specificity” were clustered with answers using the same words in the correct order.

In this case, a setter familiar with clustering in marking would have set the question as two separate “leaves”. However, for language translation exercises, word order within a sentence is critical. Thus in a diagnostic test in Italian (IT1200a Q.4.7), the answer “le abbiamo incontrato” received one mark and “abbiamo l’incontrato” none.

Conclusion

Experiments comparing relatively small differences in similarity metrics and clustering algorithms have so far proved inconclusive, yielding only small differences in the correlation of clustering results to human marking judgements. We expect further experiments with a wider range of language engineering techniques to improve performance, especially for slightly longer text answers.

Differences in types of question had much larger effects. Although clustering is most effective on very short answers, this is far from the whole story. Answers where word order is significant, or where original examples are required, for instance, need treating differently from ones where this is not the case.

Clustering is a good tool for thinking about the nature of questions and answers as well as improving speed and consistency of marking in some cases. It clearly has great potential for reducing the workload, and hence improving the timeliness, involved in formative feedback. The examples shown in this paper support our general view that fully automatic summative assessment of constructed answers is generally unsafe in view of What Students Really Say.

Short text answer questions do have paedogogic value if used thoughtfully, and are amenable to light-weight processing in an HCC framework. Analysing answer data (especially marked answer data) can bring some surprising insights into paedogogic aspects of seemingly simple questions.

References

Bloom, B., M. Englehart, E. Furst, W. Hill, & D. Krathwohl (1956) **Taxonomy of educational objectives: the classification of educational goals. Handbook I: Cognitive Domain**. New York: Longmans.

Buckley, C. & A. Lewitt (1985) *Optimization of inverted vector searches*. Proc. 8th Annual SIGIR Conference on Research and Development in Information Retrieval, pp. 97-110.

Carlson, A. & S. Tanimoto (2005) *Text Classification Rule Induction in the Presence of Domain-Specific Expression Forms*. Mixed Language Explanations in Learning Environments (XLANG), AIED (Artificial Intelligence in Education) 2005), Amsterdam.

Heyer, L.J., S. Kruglyak, & S. Yooseph (1999) *Exploring Expression Data: Identification and Analysis of Coexpressed Genes*. **Genome Research** 9:1106-1115.

Jain, A.K., M. N. Murty, & P.J. Flynn (1999) Data *Clustering: A Review*. **ACM Computing Surveys** 31:3

Lütticke, R. (2005) *Graphic and NLP Based Assessment of Knowledge about Semantic Networks*. XLANG.

Porter, M.F. (1980). *An algorithm for suffix stripping.* **Program** 14(3), 130-137

Pulman, S. & J.Z. Sukkarieh (2005) *Automatic Short Answer Marking*. Proceedings of Association for Computational Linguistics.

Salton, G. (1971) *The SMART Retrieval System: Experiments in Automatic Document Processing.* Prentice Hall, Englewood Cliffs, NJ.

Sargeant J., M.M. Wood & S.M. Anderson (2004) *A human-computer collaborative approach to the marking of free text answers*. 8th International Conference on Computer Aided Assessment, Loughborough University, Loughborough, UK, pp.361-370.

Tselonis, C., J. Sargeant & M.M. Wood (2005) *Diagram matching for human-computer collaborative assessment*. 9th International Conference on CAA, Loughborough, UK. pp. 441-456.

Wiemer-Hastings, P., E. Arnott, & D. Allbritton (2005) *Initial Results and Mixed Directions for Research Methods Tutor*. XLANG.

Wood, M.M., J. Sargeant & C. Jones (2005) *What Students Really Say*. 9th International Conference on CAA, Loughborough, UK. pp. 317-327

Zamir, O., O. Etzioni, O. Madani, & R.M. Karp (1997) *Fast and intuitive clustering of web documents*. KDD-97 pp. 287-290.

Appendix

		Average Linkage, Cluster to 50%					
		CS141203		CS141204		CS141205	
	Q1.1a	Q1.1c	Q1.3d	Q1.1a	Q1.1	Q1.2	Q3.1a
No. of Answers	153	151	137	116	93	91	27
No. of Terms	115	130	386	119	86	510	19
No. of Clusters	76	75	68	58	46	45	13
No. of Outliers	55	70	49	45	26	28	12
Avg Within Cluster Similarity	0.9695	1.0000	0.9182	0.9839	0.9455	0.8625	1.0000
% Marking Reduction	50%	50%	50%	50%	51%	51%	52%
Avg SD of Marks	0.4450	0.0000	0.2100	0.1295	0.3160	0.6340	0.0000
% Anomalous Marks	8%	0%	6%	4%	11%	22%	0%

Table 1. Cluster analysis of answers to questions across three years of the Artificial Intelligence Fundamentals course CS1412. Clusters were created using an Agglomerative Hierarchical algorithm with an Average Linkage metric used to measure distance between Clusters. In each case the algorithm was run to create a number of clusters equal to 50% of the number of answers.

Number of Answers is the total number of answers in the set and *Number of Terms* is the number of terms (words) in the Term-by-Document Matrix. This provides a measure of how variable or diverse the answers are.

Number of Clusters is the total number of clusters at the termination point while the *Number of Outliers* is the number that contain just one answer.

The *Average Within Cluster Similarity* is a measure of how similar answers are within a cluster, i.e. the average number of terms which documents in a cluster share.

% Marking Reduction indicates how much clustering has reduced the number of individual answers a human marker would have to see if they trusted the clustering completely. Whether such trust would be justified is indicated by the

Average SD of marks within Clusters, the overall standard deviation between marks within each cluster, an indication of how well the clustering correlates with the actual human marking.

The Number of Anomalous Marked Answers is another measure of that correlation, the number of answers that were not awarded the same mark as the others within a cluster, while *% Anomalous Marked Answers* gives the same value corrected for the overall number of answers in the cluster.

CS141203 Q1.1a: Here are three rules I might use in deciding how to get to work in the morning:

1. IF weather fine THEN take train
2. IF weather fine AND cold THEN take train and wear woolly hat
3. IF train drivers on strike THEN take bus

What conflict resolution strategy would you use to force rule 2 to fire? What strategy would you use to force rule 3 to fire?

Model answer: Rule 2 - specificity
Rule 3 – priority

CS141203 Q1.1c What are the three components of a production system?

Model answer: Working memory
Rule memory
Interpreter

CS141203 Q1.3d: In artificial intelligence, what is the "Turing test"?

Model answer: A simple test for "intelligence". A tester has to distinguish between communication with a human and with a machine. If they cannot tell the difference, or think the machine is a human, then the machine has passed the test

CS141204 Q.1.1a: In the "Hector's World" lab, conflict resolution is handled by "salience". Name two other conflict resolution strategies which can be used in production systems.

Model answer: Any two of rule ordering, specificity, recency, random.
NB priority is not acceptable, as it is a synonym for salience.

CS141205 Q1.1: A traditional cricket ball is red. Express this fact as a very simple semantic network, in two different ways.

Model answer: cricket ball --<has property>-- colour --<value>-- red
cricket ball --<has colour>-- red

CS141205 Q1.2: A cricket ball used in a day-night match is white. What problem does this cause for semantic networks?
Give another example of the same problem (but **not** the example featured in the lectures).

Model answer: Exceptions to default inheritance.
Anything sensible except penguins as non-flying birds, since that was the primary lecture example

CS141205 Q3.1a: The CS1412 "Hector's World" lab uses the programming environment JESS. What does "JESS" stand for?

Model answer: Java Expert System Shell

OVERVIEW OF JISC ASSESSMENT ACTIVITIES

Lou McGill

Overview of JISC Assessment Activities

Lou McGill
Programme Director
JISC
L.McGill@jisc.ac.uk

The Joint Information Systems Committee (JISC) are delighted to be one of this year's sponsors of the 10th International Computer Assisted Assessment Conference.

The immense possibilities offered by e-assessment are recognised in all sectors of education and supported by the Department for Education and Skills (DfES) as fundamental to the future success of learners. JISC recognised the importance of e-assessment for the UK education and research as long ago as the late 1990s as a part of our groundbreaking work on Managed Learning Environments. JISC realise that they have an important role to play in representing the needs of the education and research communities in this fast moving and dynamic area. As more and more software suppliers and developers become involved in producing e-assessment products so JISC are bringing the issues associated with this increasingly complex area to the attention of the communities that we serve. As more and more institutions as well as individual teachers and lecturers are beginning to use e assessment as a part of their daily teaching and learning activities so the need to provide key support such as the e-assessment glossary and roadmap becomes increasingly important. As well as providing support we have an ongoing commitment as a part of our e-learning programme to explore and research future developments and trends in this fast growing area. Last year we brought to you a proposed programme of work and this year we are pleased to bring you the fruits of this so far including:

- The launch of an online interactive e-assessment glossary
- An e-assessment roadmap
- Case studies of e-assessment
- An update on the Framework Reference Model for Assessment (FREMA) Project
- Information about our forthcoming Toolkits
- News on the important work of the JISC services CETIS and NETSKILLS
- An update on future JISC activities in this area
- A chance to meet the people involved in this work at JISC and CETIS and NETSKILLS

JISC recognise the need to work closely with our partners and associates due to our UK wide remit and are uniquely placed to provide advice guidance and research that is relevant and timely and impartial.

ESSAY EXAMS AND TABLET COMPUTERS – TRYING TO MAKE THE PILL MORE PALATABLE

Nora Mogey and Greg Sarab

Essay Exams and Tablet Computers – Trying to Make the Pill More Palatable

Nora Mogey
Media & Learning Technology Service
University of Edinburgh
Tel 0131 651 6163
Nora.Mogey@ed.ac.uk

Greg Sarab
CEO, Extegrity Inc.
Tel (USA) 415-255-2842
fax 801-912-2296
greg@extegrity.com

Abstract

Most students now complete most assignments using a computer. Word processing is standard. Yet when it comes to the end of the semester we still require most students to handwrite final examinations. Surely we can no longer claim this is an authentic assessment strategy?

At The University of Edinburgh we have been conducting trials to explore the potential for using computers in traditional examination settings. In itself, the concept is not unusual, as nearly all US law schools have been leveraging student-owned laptops on academic examinations for many years.

The additional feature we sought and have tested is the ability to sketch a diagram and include that with the text of the essay. We will briefly demonstrate the software and discuss evaluation results.

Student reaction has, predictably, been positive, but with some concerns and reservations in using the hardware/software and on issues of equity and fairness. Some found the very concept of including a diagram in an essay startling, while others thought it natural and desirable. All found it physically awkward to manipulate the tablet PC between use of the keyboard and the touch screen. Some expressed concerns about whether those students who can touch type are unduly favoured, and whether in fact this widens unfairly the inevitable inequalities between individuals, and their comments suggest it is necessary to consider differences in examiners expectations and decisions when presented with typed rather than hand written scripts. Most importantly,

support from the student body to continue to develop this approach is strong and consistent. It will not be suitable for all examinations in all subjects, but clearly this will be a useful tool for a wide variety of contexts.

Introduction

"The death of handwriting" was a recent eye-catching headline in a national newspaper (Jeffries, 2006). The article argued that although writing as a skill is valued and positively encouraged by government initiatives such as the handwriting element in the national curriculum, there is evidence children are not developing the early motor skills needed for fluent handwriting. Instead our young people are becoming "digital natives" (Prensky, 2001). As far back as 2002, 98% of UK children aged 5-18 used computers regularly (National Statistics Office). Increasingly we can expect our students to arrive at university with excellent technology skills, personally-owned equipment, and strong expectations that university will be a technologically advanced environment (Haywood et al 2004).

Questions can be raised about examinations and the contexts for which they are or are not an appropriate assessment tool (Rowntree, 1977; Howell, 2003; Harris, 2005). The present study assumes essay examinations will continue to feature in the assessment portfolio for some time and explores a primary method to introduce computers into that setting.

US law schools most commonly assess students via a single, high-pressure, 3-hour essay per course, and examinations have long been held on computer (Augustine-Adams et al 2001). Students typically provide their own laptops (a small number of school-owned computers are available at a few schools), and are responsible for installing and operating special exam software. This offers advantages of student familiarity with, and responsibility for, the machine used for testing. Test questions are normally distributed on paper, further maintaining the familiar traditional environment.

An essay in law, in common with many humanities subjects, will typically be largely text. This is not the case in all disciplines, especially the sciences, where it is usual to wish to include diagrams, sketches, graphs and the like in an essay response. The notion of using a tablet PC which could be used either to enter text or sketch a diagram seemed attractive. An exploratory project was successfully established as part of the Change Academy 2004.

Software Selection and Development

Initial investigations identified several pieces of software concerned with the collection of responses in an examination setting. The most pertinent tools provided a secure typing environment where no other applications could run concurrently, offered appropriate data encryption, and carefully saved and protected student work. We immediately discovered the security aspect was so effective that it also blocked the very tablet functionality we were keen to exploit. Upon explaining this problem to a number of vendors, Extegrity Inc.

agreed to collaborate with us to facilitate the inclusion of figures or sketches with a typed exam.

Evaluation

15 student volunteers participated in the evaluation held in January 2006.

The afternoon comprised a short overview of the project, time to practice using tablet computers owned by the University running Extegrity's specially enhanced Exam4 software, and a 1-hour written "exam". Students were observed during the exam and feedback was sought on paper and via two focus groups.

The "exam" was modelled on a traditional paper-based exam, with a printed question provided. Candidates launched the software and completed the initial administrative procedure up to a "Wait" screen. The invigilator verbally confirmed all students had successfully reached that screen, then invited them to turn over the exam question and proceed.

At the end of the exam the invigilator instructed the candidates to stop and follow the software's exit procedure. In this case student responses were submitted via USB flash drive but it could equally well be to a connected network drive or any other media desired. The saved files are encrypted and cannot be opened without a security key.

The student volunteers were mostly active members of the Edinburgh University Students' Association, although at least 2 were not. 6 were male and 9 female (of which 2 were mature students) representing 10 schools from 2 of our 3 colleges. Although this session was not advertised as being about use of computers, the group were highly computer literate. 3 had previously taken an examination using a computer.

Student Feedback

Observation suggested that despite the lack of practice students had few problems with the overall process. A clear recurring concern was that rotating the screens of the tablet PCs was time consuming, physically awkward and distracting both to others and one's own train of thought. Some additional functions were requested in the software (e.g.: bullet points, tables) and there was a desire to have the images embedded within the main body of the text. However, no one was worried about whether their work had been saved correctly.

The students were open to the suggestion of a sensible role for computers in essay exams. While broadly supportive of exploring this idea, perhaps even expressing a small amount of enthusiasm, there were also very strongly expressed reservations. Many were concerned about the impact of

differences in typing abilities, and all stressed that sufficient practice time would be critical. Discipline differences were evident, with mild confusion being expressed by some students as to why anyone would ever want to include a drawing in an exam. An easy to implement suggestion/request was to provide scrap paper for those who wished to use it.

Members of both focus groups stressed they had concerns about possible unfairness due to differences in typing skills, closely associated with concerns about how a typed exam might be marked differently to a handwritten exam, and that students would feel they had to go back and correct typing errors which they would just leave in a handwritten submission. Students did not view this unfairness as being equal or equivalent to any which is inherent to the current system with handwritten examinations.

There was no clear and consistent feeling about whether students would do better or worse on such an examination: some students welcomed the idea and others had significant concerns.

Conclusion and Future Directions

Despite broad encouragement from our student evaluators to continue this study there are some practical difficulties, particularly in accommodating drawing capabilities. Few students own tablet PCs, and it would be expensive to purchase a set large enough for a typical first- or second-year class.

Since so many students do own non-tablet laptops, a promising option is the provision or requirement of inexpensive peripheral USB tablet devices for exams. Other less favourable options include: proceeding only with "non-drawing" disciplines; allowing diagrams drawn on paper to be submitted with the examination script; and/or, restricting use of tablets to contexts where diagrams are integral and where student numbers match the resources available.

In the longer term it will be necessary to adopt invigilation procedures and general protocols for conducting this type of assessment such as already established in institutions where objective testing is well-embedded (ex:Uni Dundee).

It is recognised that further study is needed regarding the more psychological issues raised by the students (expectations about differences in marking handwritten scripts versus typed scripts; how students would actually spend the precious examination time if using a keyboard) before we could proceed to widespread adoption of this method of assessment. In this we can draw upon the experiences of US law schools and students, and Extegrity, veterans of hundreds of thousands of computer exams.

Whether traditional examinations are the future may be questionable.

Nevertheless this early test has been encouraging and well-received by our students, demonstrating it is possible to mix new technologies with old

assessment methods and perhaps make the bitter pill of examinations a little easier to swallow.

Acknowledgements

Initial planning for this project was undertaken with the support of the Change Academy.

The support and assistance of the Edinburgh University Students' Association and the individual students who participated in this evaluation was critical, and continues to be much appreciated.

References

Augustine-Adams, K. Hendrix, S.B. and Rasband, J.R. (2001) Pen or Printer : can students afford to handwrite their exams? *Journal of Legal Education,* 51,1, pp118-129

Harris, R. (2005) Testing Times: Traditional Examinations and Asynchronous Learning. *Journal of Geography in Higher Education*, 29, 1, pp 101-114

Haywood, J. et al (2004) The Student view of ICT in Education at The University of Edinburgh: skills attitudes and expectations. *ALT-C 2004 Proceedings*

Howell, S. L. (2003) e-Learning and paper testing: why the gap? *Educause quarterl*y, 4.

Jeffries, S. (2006) The death of handwriting, *The Guardian* 14 Feb 2006

National Statistics Office http://www.statistics.gov.uk/ (Accessed May 2006)

Prensky, M. (2001). Digital natives, digital immigrants. *On the Horizon, 9*(5), 1–2.Available:www.marcprensky.com/writing/Prensky%20-%20Digital%20Natives,%20Digital%20Immigrants%20-%20Part1.pdf (accessed May 2006)

Rowntree, D. (1977) How to Assess? In *Assessing students: How Shall We Know Them?* Kogan Page

University of Dundee online assessment procedures http://www.somis.dundee.ac.uk/academic/caapolicy.htm (accessed May 2006)

Change Academy http://www.heacademy.ac.uk/923.htm

http://www.softwaresecure.com/pdf/PaperTests.pdf

DEVELOPMENT OF WEB BROWSING TECHNIQUES TO CAPTURE RESPONSES IN THE CONTEXT OF ENGLISH LANGUAGE SKILLS ASSESSMENT

Jenifer Moody and Jeremy Swift

Development of Web Browsing Techniques to Capture Responses in the Context of English Language Skills Assessment

Jenifer Moody and Jeremy Swift
Education Development International plc.
International House
Siskin Parkway East
Middlemarch Business Park
Coventry
CV3 4PE
+44(0)24 76516 511
jenifermoody@ediplc.com

Abstract

The assessment of speaking skills in English as a foreign language presents pedagogical and logistical challenges, which are being exacerbated by the increasing demand for British qualifications from students based overseas.

To address these challenges, and to reduce the time lapse between taking the test and issuing the results, we have developed an on-line version of a traditional high-stakes Speaking test, using Real Time Messaging Protocol (RTMP).

Introduction

According to figures provided by the British Council[1] English has an official or specialist status in at least 75 countries. It is spoken as a second language by 375 million people, as a foreign language by a further 750 million people and to 'some extent' by a quarter of the world population.

It is the official language of the European Central Bank, of the United Nations and of maritime communication and of international air traffic control

The process of globalisation[2] is increasing the dependence on English as the (primary) means of communication in business and industry. This is producing a concomitant increase in the number of people wishing to learn English[3] and to have their competence certified by means of qualifications.

This world-wide and increasing interest in communicating in English presents pedagogical and technical challenges for the assessment industry.

Pathways to Proficiency[4] produced sets of scales for the four language modes – listening, speaking, reading and writing. This paper will focus on the challenges related to assessing Speaking.

Speaking - The Pedagogical Challenges

The requirements for demonstration of competence in speaking the English language are demanding at even quite low levels. At the equivalent of Level 1 (Adult Literacy) (defined as B2 Vantage of the Common European Framework) an individual is required to be able to[5]

- *Give clear, systematically developed descriptions and presentations on a wide range of subjects …….expanding and supporting ideas with subsidiary points and examples.*

- *Participate actively in routine and non-routine formal discussion*

- *Contribute, account for and sustain his/her opinion, evaluate alternative proposals and make and respond to hypotheses.*

To develop an assessment against these criteria that will produce the necessary levels of validity and reliability is a challenge.

Some Awarding Bodies[6] assess Speaking skills through interviews with visiting assessors. This method has the merit of offering both face and construct validity, but is administratively difficult for centres and students. It is also time-consuming, expensive and unreliable. It is therefore not a suitable method for assessing large numbers.

In other contexts[7] speaking capability is assessed by means of face-to-face interviews that are recorded and then sent elsewhere to be marked. Whilst the use of traditional recording techniques (typically cassette tapes) overcomes some of the cost and administration problems associated with personal interviews, the method is still time-consuming and not particularly reliable.

In yet other contexts, Speaking is not assessed as a separate skill at all. Eckstein and Noah[8] (1993) point out that in neither China nor Japan is there any attempt to assess oral skills as a part of school level examinations in English. In both countries, multiple-choice is the predominant assessment format for the other components. Here the emphasis is on reliability, but at the expense of a valid assessment of speaking capability.

To meet the increasing demand from globalisation and at the same time present students with the opportunity of an appropriate assessment of their capabilities there needs to be some way of combining the advantages of the face-to-face personally conducted interview with the opportunities offered by new technologies for more efficient and more reliable assessment.

The key to this is the development of a secure, web-based system for the delivery of an authentic assessment, combined with the creation of a distributed on-line marking facility.

The Project

The project was to take an existing speaking test for an international high stakes English qualification[9], and remove the need for recording student responses on cassette tapes by using Real Time Messaging Protocols (RTMP) and streaming student answers.

The project also required the development of a distributed, on-line marking facility that allowed markers to access tests for marking from anywhere in the world, and which also included functions that would enhance the reliability of marking.

The Technical Solution

There were two principal challenges that needed to be overcome:

- The need for a reliable connection between the test delivery interface and the web application server;
- The security implications of the networks and firewalls at the centres running the tests.

Macromedia Flash components provide a development infrastructure that enables connections to remote services that are exposed by application server developers and web services. Macromedia Flash Remoting simplifies the application development process by providing us with a programming mode and runtime support for connecting the application directly to remote server objects

Using Macromedia Flash Remoting, we can easily connect ActionScript client logic directly to our remote services without writing any wrapper code, proxy code, or data marshalling code. Macromedia Flash Remoting exposes well-defined application APIs and services (whether implemented in C#, Java, or ColdFusion) transparently to Macromedia Flash as ActionScript APIs. Macromedia Flash Remoting also adds a rich debugging capability and a service browser between the Macromedia Flash client and the server, providing us with an optimized development experience in creating Rich Internet Applications using Macromedia Flash

Macromedia Flash Communication Server provides the same capabilities as Macromedia Flash Remoting except that the Flash Communication Server communicates with the application server instead of the Flash movie. The Flash movie communicates with the Flash Communication Server via the real-time RTMP (Real-Time Messaging Protocol) protocol for audio/video/messaging applications.

The ELSA Speaking Test uses the Macromedia Flash Communication Server for the streaming of candidate answers. By default this technology uses RTMP (Real-Time Messaging Protocol).One hurdle we had to overcome was

the security implications of networks and firewalls at the centres running the tests. It became apparent that not all of the centres had the same security and firewall settings enabled to allow successful connections, to overcome this we built an online diagnostic tool that centres use which provides statistical information on the available open ports required to allow successful RTMP traffic. Armed with this information we can assist the centres in making a valid data stream connection to and from the Flash Communication Server.

Evaluation

Technical

To date, the system has been successfully piloted in the Middle East and South America as well as Europe.

Technically, it works well, although experience has shown that it is necessary to engage in quite extensive dialogue with centres in advance to ensure that their infrastructure will provide a suitable testing environment.

Reliability

The project has identified a number of issues in relation to the marking of Speaking assessments. So far, the distributed marking system has been used with only a small number of examiners.

The main advantage has been a great reduction in the period of time taken to mark the tests and return the results. A turnaround time of five days (from date of test to time of return of results) is now being achieved regularly. This is proving to be of great benefit to centres and students who are using the test as a pre-course filter.

We have as yet insufficient evidence to comment on any effects on reliability in marking. Using the system, markers are able to compare extracts from student responses with each other and with exemplars. As they consider each question, a simple slider enables them to compare the level descriptors with the responses.

Feedback from markers has been positive, and we will carry out a more detailed evaluation once more markers are involved.

Student Experience

The evidence relating to the effects of screen-based testing on student achievement is mixed and appears to be context dependent. Comparisons of paper and computer based versions of psychological tests show equivalence[10] but other studies in educational contexts suggest different results[11].

Student feedback has been positive, with most reporting that they found the experience of taking a speaking test on-line less stressful than using a tape recorder.

As yet, there has been no opportunity to compare paper-based and on-screen results, although it is planned to do this once the volume of on-screen students increases.

Conclusion

The development of a secure web-based system for providing valid speaking tests looks likely to increase reliability. Future developments will look at further enhancing the authenticity of the assessments, and consider the implications for the inclusion of more realistic settings.

References

British Council at www.britishcouncil.org/english/engfaqs.htm

'Globalisation' is an imprecise term, covering many aspects of politics, economics and cultures. Its origins are complex, and probably not all that recent. It is currently used as a short hand way of describing the trends towards interdependence between countries encouraged by the rapid developments in information technology,

See, for example, the figures in the **Future of English***,* the report on the British Council's English 2000 project.

Pathways to Proficiency: The Alignment of Language Proficiency Scales for assessing competence in English Language, DfES, London, 2003

DfES (2003) **Pathways to Proficiency, The Alignment of Language Proficiency Scales for assessing competence in English Language,** DfES, London, 2003, Map 4, Speaking Scales, pg 40.

For example, the Assessment and Qualifications Alliance (AQA) and Jet Set and LCCI International Qualifications with SEEFIC.

GCSE and A level examinations, for example.

Eckstein, A., and Noah, H., (1993) **Secondary School Examinations**: **International Perspectives on Policy and Practice,** New Haven, Yale University Press

The project used the English Language Skills Assessment (ELSA) offered by LCCI International Qualifications. The assessment consists of four components, (Reading, Writing, Speaking and Listening) and covers levels A1-C2 of the Common European Framework. There are CBA and paper versions of all components. In traditional mode the Speaking test is recorded in the centre and tapes returned to the UK. These are then forwarded to examiners for marking.

McFarlane, A., (2003) Editorial. Assessment for the Digital Age, in **Assessment in Education,** Vol. 10 (3), pg 263

11. For example, Johnson, M., and Green, S., (2004) On-Line assessment: the impact of mode on student performance*,* paper presented to the **British Educational Research Association Annual Conference**, September 2004, Manchester

ASSESSMENT FOR LEARNER SELF-REGULATION: ENHANCING THE FIRST YEAR EXPERIENCE USING LEARNING TECHNOLOGIES

David Nicol

Assessment for Learner Self-regulation: Enhancing the First Year Experience Using Learning Technologies

David Nicol
Centre for Academic Practice and Learning Enhancement
University of Strathclyde
Scotland
d.j.nicol@strath.ac.uk

Abstract

Concerns about non-completion and the quality of the first year student experience have been linked to recent changes in higher education such as modularisation, increased class sizes, greater diversity in the student intake and reduced resources. Improving formative assessment and feedback processes is seen as one way of addressing academic failure and of enhancing the learning experience and students' chances of success in the early years of study. This paper argues that if this is to happen a broader perspective on the purposes of formative assessment and feedback is required, one that links these processes to the development of learner self-regulation. It then shows, through two case studies, drawn from the Re-engineering Assessment Practices (REAP) project, how ICT might support formative assessment processes and the development of self-regulation in large first year classes. Finally, the paper presents a set of principles for the effective design and evaluation of formative assessment and feedback processes in relation to learner self-regulation.

Introduction

Across the higher education sector there is a growing interest in the quality of student learning experience in the first years of undergraduate study. This interest is fuelled by statistics showing poor course non-completion rates and by a recognition that the first year lays the foundation for learning in later years. Yorke and Longden (2004) in studying retention issues across a number of countries have identified four broad reasons why students leave academic programmes (i) flawed decision making in initial choices (ii) events that impact on students' lives outside the institution (iii) students' experiences of the programme and the institution and (iv) failure to cope with the academic demands of programmes. This paper is primarily concerned with the last two reasons: it explores how formative assessment practices might be used to enrich the first year experience and enable students to develop their capacity for self-regulated learning. It also explores how information and communication technologies (ICT) might support formative assessment

practices. Case study applications, drawn from a large-scale re-engineering assessment project led by the University of Strathclyde, are used to illustrate some possibilities. A key idea in the retention and non-completion research is the need to maximise students' sense of, and chances of, success particularly when they enter HE and in the early years of study. The concepts of self-regulated learning and academic success are central to this paper.

Formative assessment and academic failure

There is a considerable body of evidence showing that the number of opportunities available for formative assessment and feedback is an important variable in non-completion by students in the early years of study, even though a direct causal connection has been difficult to prove (Yorke, 1999). Yorke (2004) has argued that where students are uncertain about their ability to succeed formative assessment and feedback is of particular significance. However, over the last 10 years, modularisation, larger student numbers in first year classes, greater diversity and reduced staff-student ratios have all had a negative effect on formative assessment practices. These negative effects include fewer opportunities for students to clarify what is expected of them, a reduction in feedback on assignments and in class, and an increased emphasis on summative assessment at the expense of formative assessment (Yorke and Longden, 2004). The latter has resulted in an excessive concentration by students on getting good marks and playing the assessment game rather than focusing their effort on deep and lasting learning. These changes have also been shown to impact on the students' sense of self and on their motivation and self-confidence.

How might assessment practices change in order to enhance the first year experience and increase students' chances of success? A recent literature review carried out by Gibbs and Simpson (2004) was directed at addressing this question. They examined a wide range of case studies and were able to identify eleven conditions under which assessment might support student learning and increase the likelihood of academic success. The conceptual framework underpinning these conditions (and an associated assessment experience questionnaire) was based on two over-riding principles. The first principle, which draws on Chickering and Gamson's (1987) research, is that assessment tasks should be designed to ensure that students spend their study time in productive ways: tasks should encourage 'time on task' (e.g. in and outside class), should lead to a more even distribution of study effort (over the timeline of the course), should engage students in deep rather than surface learning and should communicate clear and high expectations. The second principle is about the effective provision of feedback to students on their academic work: feedback should be of sufficient quantity, timely, of high quality and actually used by students to make improvements in their work.

Although Gibbs and Simpson (2004) offer sound advice for anyone wishing to improve formative assessment, their eleven conditions are largely teacher-driven. It is the teachers who are expected to ensure that students spend time on task and that they receive appropriate feedback. While what the teacher does is an important determiner of academic success there are other

perspectives. For example, Yorke and Longden's (2004) argue that a key component of academic motivation and success is that students perceive themselves as agents of their own learning. Indeed, these researchers maintain that the student perspective is the gateway to solving what they call the 'retention puzzle'. If students are to have a sense of control over their own learning then formative assessment practices must also help them develop the skills needed to monitor, judge and manage their learning. In line with this approach, the conceptual model underpinning formative assessment practices in this paper is based on developing learner self-regulation (see, Nicol and Macfarlane-Dick, 2006).

Alongside the need to rethink the purposes of formative assessment there is also a need to rethink the methods by which formative assessment is delivered. Recent advances in information and communication technologies (ICT) are having a large impact on the organisation and delivery of student learning in HE. There is also a growing interest in the use of computers to streamline the delivery of formative assessment tests and of teacher feedback. While the implementation of some of Gibbs and Simpson's eleven conditions could be supported using computer-assisted assessment (e.g. the provision of rapid feedback through online tests), Gibbs (2006) is less convinced about the value of CAA. He maintains that:

There is very little evidence that the increase in the use of computer-based assessment has a beneficial impact on the quality of student learning, though there is some evidence that it has increased its quantity. [Gibbs, p18]

This paper demonstrates ways in which ICT can be used to support the development of learner self-regulation and the delivery of teacher feedback.

Self-regulation and Student Success

Formative assessment is defined in this paper as 'assessment that is specifically intended to provide feedback on performance to improve and accelerate learning'. (Sadler, 1998, p77). Academics tend to think of formative assessment in terms of the judgements they make about students' academic work and the provision of feedback. However, this paper takes a broader view of the source of formative assessment. It is especially concerned with involving students in evaluative judgements about their own work and the work of their peers. The ability to monitor, critically assess and correct one's own work is a key goal of higher education and of lifelong learning.

In 2006, Nicol and Macfarlane-Dick reinterpreted the literature on formative assessment and feedback in relation to learner self-regulation. From this they were able to identify seven principles of good feedback practice that if implemented would contribute to the development autonomy in learning. Each of these principles is defined in detail in the earlier paper with the supporting research and example their implementation. Table 1 presents the seven principles.

Good feedback:

1. helps clarify what good performance is (goals, criteria, standards)
2. facilitates the development of self-assessment and reflection in learning
3. delivers high quality information to students about their learning
4. encourages teacher and peer dialogue around learning
5. encourages positive motivational beliefs and self esteem
6. provides opportunities to close the gap between current and desired peformance
7. provides information to teachers that can be used to help shape teaching.

Nicol and Macfarlane-Dick (2006)

Table 1: Seven Principles of Good Feedback Practice

The work of Nicol and Macfarlane-Dick builds on that of other researchers who have emphasised the importance of developing autonomy in both learning and assessment processes (e.g. Knight and Yorke, 2003; Boud, 2000). However, it departs from the work of others in one important respect. In the model it is assumed that students are always engaged in self-regulation but that some students are better at self-regulation than others; and it is the weaker students that need opportunities to enhance their sense of control (Nicol and Macfarlane-Dick, 2006). There are at least three reasons for this argument. Firstly, students are always informally engaged in the self-regulation of learning when they engage in academic tasks (e.g. writing an essay). Indeed self-regulation is logically implied by active and constructivist thinking (Winne, 2006). In constructing meaning students are already assumed to be active agents of their own learning.

Secondly, when students receive feedback from teachers they must engage in self-assessment if they are to use that information to improve academic performance: that is, they must decode the feedback message, internalise it and use it to make judgements about and modify their own work. This implies that self-assessment is at the heart of formative feedback (from teachers) and is a key component of self-regulation. Thirdly, students in some large first year classes in higher education (e.g. over 500 students) receive almost no feedback and still make progress. Hence they must be making ongoing judgements about, and managing aspects of, their own learning - otherwise they would not be able to make progress. In summary if students are already involved in self-assessment and self-regulation then the argument is that higher education teachers should build on this capacity rather than focus all their efforts on providing expert feedback.

The REAP Project

The following sections present two case studies showing how ICT can support the development of learner self-regulation. Also provided are some illustrative examples of how learner self-regulation might be supported using multiple-choice tests. Each of the case examples uses different technologies (discussion board, electronic voting systems, and multiple choice tests). The

context of these case studies is the Re-engineering Assessment Practices [REAP] project, one of six projects funded by the Scottish Funding Council under its e-learning transformation initiative.

The overall aim of the REAP project is to demonstrate learning quality enhancement and more effective use of staff time in large first year classes (150-800 students) through the application of learning technologies. The project involves three Scottish HE Institutions each piloting different approaches and technologies across a range of disciplines. The REAP project draws on the Nicol and Milligan, 2006 research in that a key objective of assessment re-engineering is to lay a foundation for autonomy and self-regulation in learning during the first year.

Example 1: Psychology

The first year Basic Psychology course is designed to introduce all students to key findings, theories, and debates in general contemporary psychology. In addition the class provides continuing students with an introduction to a number of specific areas of study within psychology which are dealt with in depth in second, third, and fourth year classes. The course comprises six topic areas delivered by 48 lectures, 4 tutorials and 12 practical laboratories over the year. The class size is approximately 550 students. Before the changes reported here assessment comprised two paper-based multiple-choice tests over the year (25%), tutorials (4%), participation in an experiment (5%) and a final exam where students write 3 essays from eight (66%). Feedback was only available through marks given on the multiple-choice tests and there were concerns that students were not given any feedback on their writing, essential for good exam performance. Technology-supported assessment was seen by the class leader as having the potential to enhance the first year experience, increase students' understanding of the topics being studied and enhance success in written work without increasing staff workload

The Pilot Study

In the psychology pilot, the basic class was re-designed to provide opportunities constructive formative assessment (scaffolding) linked to supportive peer discussion. This project draws on research showing cognitive gains where peer discussion is directed at the resolution of conflicting views. The discussion board within the institutional virtual learning environment (WebCT) is the technology in use.

Seventy-eight students were invited to participate in the pilot study (15% of class). The students were divided into groups with a maximum of six students per group. There was an initial induction task where students were asked to introduce themselves to each other within their groups via the online discussion board. The main academic task followed this and involved students being presented with three questions of increasing complexity in a specific topic area (e.g. human memory) over a number of weeks. For the first question they were asked to post an individual 50 word response to a

private submission area in WebCT: this response could not be seen by other students. They were then directed to engage in an online discussion about their answer; the instructions were to debate/argue what they believed the correct answer to be. For the second question they are asked to engage in online discussion in their groups and to post an agreed 100 word response to the discussion board by a certain date. For the third question they also engaged in online discussion but posted a 300 word response. Before students engaged with the second and third questions they were directed to a model answer written by the teacher; they could also retrieve a model answer after the 300 word response.

Relation to seven feedback principles

Key features of this pilot are that the task questions are progressively more difficult, that responses move from an individual to a group that there is a model answer for comparison at each stage. Tutors provide no feedback neither do they moderate the discussion. What is important here however is how this course design implements the seven principles of good feedback and helps develop learner self-regulation:

- Providing a model answer is one way of progressively clarifying expectations and helps clarify what good performance is (Principle 1)

- Students engage in self-assessment (reflection) by comparing their own responses against the model answers (Principle 2)

- There is online peer discussion around the learning task with the goal of reaching consensus about each group's submitted responses (Principle 4)

- The increasing complexity of the questions scaffolds and stages learning development and the focus on learning goals rather than marks should enhance students' motivation (Principle 5)

- The repeat and progressive nature of the task provides some opportunities to close the gap between desired and actual learning (Principle 6)

Commentary

Preliminary findings from focus groups and questionnaires show that that the students were positive about this learning experience. They reported that working collaboratively has enhanced their understanding of the discussion topic (92%). Typical student comments were "*we know everything there is to know about this topic now*" and "*I found it very beneficial, at the time I did not realise how much I was learning…it was learning without thinking about what I was doing*". Another finding was that the early induction task where students introduced themselves helped create more supportive social interaction in the first year as evidenced by the discussion board postings. Being part of a

large first year class does not guarantee, and may even inhibit the, establishment of social contact with others.

One question raised by the pilot is whether these peer discussion tasks should be compulsory or voluntary. Not all students participated in the online discussions and although refinements in instructions are possible this might always be the case. However, making the peer discussion compulsory would have significant implications for teachers' time as they would have to monitor contributions. An argument for leaving this task voluntary is that the feedback is an extra resource to support the first year experience; it can be used by students if they wish. This type of resource would support a movement to more flexible learning scenario.

The findings from this pilot have given the Department of Psychology the confidence to propose a radical redesign of the first year class commencing in 2006/7, abolishing half the scheduled lectures and replacing these with similar online group exercises and making self and peer feedback core components of the class. This methodology is easily transferable to other courses and is simple to implement and only involves a standard tool in any VLE (discussion board).

Example 2: Mechanical Engineering

The second example explores how a range of technologies including electronic voting systems are being used to support assessment practices and the development of learner self-regulation in mechanical engineering. Eight years ago the Department of Mechanical Engineering at the University of Strathclyde embarked on a radical change in its teaching methods for first year students (see Nicol and Boyle, 2003; Boyle and Nicol, 2003). The aim of the New Approaches to Teaching and Learning in Engineering (NATALIE) was to introduce collaborative learning in large lecture classes. The standard lecture/tutorial/laboratory format was replaced by a series of two-hour active-learning sessions involving short mini-presentations, videos, demonstrations and problem-solving all held together by peer instruction. Peer instruction is a form of Socratic Dialogue or teaching by questioning' pioneered by Mazur at Harvard (1992) using electronic voting technologies.

A typical peer instruction class would begin with the teacher giving a short explanation of a concept or presenting a video demonstrating the concept (e.g. force in mechanics). This is followed by a multiple-choice question test.(MCQ). Students respond to the concept test using handsets (similar to a TV remote) that send signals (radio frequency or infrared) to receivers linked to a computer. Software collates responses and presents a bar chart to the class showing the distribution across the alternatives. In peer instruction, if a large percentage of the class have incorrect responses the teacher instructs the class to: 'convince your neighbours that you have the right answer'. This request results in students engaging in peer discussion about the thinking and reasoning behind their answers. The learning gains from this procedure have been interpreted in terms of cognitive conflict and scaffolding both of which

have been shown to benefit learning (Nicol and Boyle, 2003). After the discussion the teacher usually retests the students' understanding of the same concept test. Another strategy is for the teacher to facilitate 'class-wide discussion' on the topic by asking students to explain the thinking behind their answers. The EVS sequence usually ends with the teacher clarifying the correct answer. There are many other ways of using EVS to facilitate interaction and collaborative and EVS have been used across a range of disciplines. In Interactive Mechanics where EVS is used, class size is 260 students (there are two sessions of 130 with each EVS class lasting two hours) summative assessment comprises 10 fortnightly written homework exercise, a two-hour class test and a written exam. .

Through REAP project funding, the Department of Mechanical Engineering is piloting new uses of EVS software (e.g. ranking tests) as well as other web-based tools such as Intelligent Homework systems. Two developments are important in relation to this paper. Firstly, the use of online tests has been integrated with the use of electronic voting. Students are presented with online MCQs before the interactive lecture sessions (EVS). The teacher then uses the results of these tests to establish areas of weakness and to determine the focus of the classroom EVS sessions. This procedure, often called 'just-in-time-teaching' (Novak et al., 1999), is a way of targeting teaching to students' needs and level of understanding. A second innovation is the use of confidence or certainty-based marking (CBM) during EVS sessions. This uses multiple-choice questions but students must rate their confidence (certainty) in their answer. This is being piloted as formative assessment using the rules in Table 2 with the intention of using this for summative assessments at a later time. CBM requires that students engage in metacognitive thinking – to step back and reflect deeply about whether there is good justification for their answer.

Degree of Certainty	Low	Medium	High	No reply
Mark if correct	1	2	3	0
Penalty if wrong	0	-2	-6	0

Table 2: Scoring regime for Certainty-based marking

Relation to the seven feedback principles

The use of EVS in Mechanical Engineering is a powerful example of an integrated implementation of the seven principles. However, for the sake of analysis we have separated out the implementation of each principle as it applies to the EVS class:

- Learning goals are clarified through iterative cycles of tutor presentation, test and re-tests of concepts using MCQs (Principle 1)

- Opportunities for self-assessment and reflection are available when the teacher provides the concept answer at the end of the EVS test

sequence and when students reflect on their answer during confidence-based marking. Reflection is also possible after the bar chart presentation of class response. (Principle 2)

- Teachers normally provide feedback during class in response to students' questions and at the end of each concept test sequence to clear up any misunderstandings. (Principle 3)

- Peer dialogue is integral to peer instruction and class-wide discussion and student-tutor dialogue occurs during class-wide discussion. (Principle 4)

- The EVS class is focused on learning goals rather than performance goals and the step-by-step progression in difficulty of the concept questions both help maintain motivation. (Principle 5)

- The continuous cycle of tests, retests and feedback ensures that students have opportunities to 'experience' a closing of the gap between desired and actual performance (Principle 6)

- A great deal of information is available to the teacher about areas of student difficulty. This is used to shape in-class teaching. The bar chart feedback also gives the teacher instant feedback about areas of difficulty and asking students to explain answers during class-wide discussion uncovers conceptual misconceptions. The information provided before class through the web-based MCQs links out of class (homework) with in-class activities: this feedback can be also inform in-class teaching (Principle 7)

Commentary

Extensive evaluations have been carried out in engineering mechanics showing significant learning gains (Nicol & Boyle, 2003; Boyle & Nicol, 2003). Overall the changes have been a huge success both in terms of student end of year performance in exams and in terms of retention. There has been a reduction from 20% non-completion to 3% the largest gain in any course within the University. Also, since the introduction of concept tests with electronic voting, attendance at class remains high throughout the year (unlike similar lecture based classes]. Further evaluations of confidence-based marking and intelligent tutoring are now being carried out.

Discussion

The two case studies reported above show how ICT can be used to support a broad range of formative assessment processes in large first year classes. A key issue in the literature on formative assessment is how to move students from being dependent on teacher feedback to being able to generate their own feedback on learning. These case studies address this issue in that they both involve elements of self assessment, peer and teacher feedback implemented in ways that support the development learner self-regulation.

But what are the potential limitations of these methods? Firstly, it should be pointed out that the Psychology study is currently in pilot mode and there is a need to scale this up to the complete student cohort of 550 and carry out a full evaluation. A second issue is the balance of learner self-regulation and teacher direction. Taking a purely self-regulated learning perspective one might argue that it is still the teacher that is directing students' learning and, in particular, their interactions with the subject matter.

In addressing this issue, it is important to note that there is considerably more autonomy built into these classes than in traditional teaching approaches. A second point, is that these are first year classes and a clear structure for learning is perhaps appropriate at this level, although this argument might not be appropriate in later years. However, it would be possible to extend learner autonomy by re-examining the case studies in the light of the seven principles. For example, one criticism of the EVS procedure might be that students are always engaged in tests formulated by the teacher. But this could be changed by having students construct tests for use in the class themselves. This would ensure that they are actively engaged in generating assessment criteria and example questions from their subject discipline (principle 1). This strategy might be more appropriate with experienced students.

One interesting observation from one of these case studies is the role played by objective multiple choice tests. Earlier in the paper attention was drawn to Gibbs' (2006) comments about the weaknesses of MCQ tests. Yet, the Mechanical Engineering example shows that it is not the test itself that is important but the context of its use. Considerable power is gained when assessment principles underpin the implementation of these tests as occurs in the EVS classroom and when the implementation includes a blend of online and offline interactions (as with just-in-time-teaching).

In the introduction, this paper also outlined Gibbs and Simpson's approach to enhancing formative assessment and feedback processes. Their concern was with the nature of the feedback provided by the teacher (its timeliness, quality, quantity and use), and that students spend their study time in productive ways. Their eleven conditions (based on these two broad principles) are important and in fact complement the seven principles advocated in this paper. Indeed, if the two case studies presented in this paper had been analysed in terms of these eleven conditions it would have been evident that many of them were satisfied.

A key outcome of the REAP project is the value of having robust formative assessment principles derived from research when thinking about the design of assessment practices. As well as being important in design such principles are also valuable in the evaluation of changes in practice. Both the Gibbs and Simpson (2004) framework and the Nicol and Macfarlane-Dick (2006) principles are a first step in this regard. Future research might see some merging of these frameworks. Indeed, this work is already underway at least in relation to written feedback (see, Brown and Glover, 2006). The

development of this research will not just help enhance the first year experience but should also benefit students in later years.

Acknowledgements

The author thanks the Scottish Funding Council for funding the REAP project as part of its e-learning transformation initiative. David Nicol would also like to thank the REAP project team, Catherine Owen, Jenny Booth and Martin Hawksey for supporting the REAP implementations reported in this paper and for advice during its preparation. He is also grateful to the two departments whose cases studies are portrayed in this article, and in particular to Jim Boyle (Mechanical Engineering) and Deidre Kelly (Psychology).

References

Boud, D. (2000), Sustainable assessment: rethinking assessment for the learning society, *Studies in Continuing Education*, 22(2), 151-167.

Boyle, J. T. and Nicol, D.J. (2003). Using classroom communication systems to support interaction and discussion in large class settings, *Association for Learning Technology Journal* [ALT-J], 11(3), 43-57.

Chickering, A.W. and Gamson, Z.F. (1987), Seven principles for good practice in undergraduate education, Wingspread Journal 9(2), special insert.

Gibbs, G and Simpson, C. (2004) Conditions under which assessment supports students' learning, Learning and Teaching in Higher Education, 1, 3-31. Online. Available at: <http://www.glos.ac.uk/adu/clt/lathe/issue1/index.cfm> (accessed 6 April 2006)

Gibbs, G. (2006), Why assessment is changing. In C. Bryan and K. Clegg (Eds), *Innovative Assessment in Higher Education*, Routledge., London.

Brown, E. and Glover, C. Evaluating written feedback. In C. Bryan and K. Clegg (Eds), *Innovative Assessment in Higher Education*, Routledge., London.

Mazur, E (1997), *Peer Instruction: a user's manual*, Prentice Hall.

Knight, P. and Yorke, M. (2003), Assessment, learning and employability. The Society for Research into Higher Education and Open University Press.

Nicol, D.J. and Boyle, J.T. (2003). Peer instruction and class-wide discussion: a comparison of two interaction methods in the wired classroom, *Studies in Higher Education*, 28(4), 477-73.

NICOL, D, J. & Macfarlane-Dick (2006), Formative assessment and self-regulated learning: A model and seven principles of good feedback practice, *Studies in Higher Education*, 31(2), 199-216

NICOL, D. J. & Milligan, C. (2006), Rethinking technology-supported assessment in terms of the seven principles of good feedback practice. In C. Bryan and K. Clegg (Eds), *Innovative Assessment in Higher Education*, Routledge. London.

Novak, G.M., Patterson, E.T., Gavrin, A.D. and Cristian, W. (1999), *Just-in-time-teaching: blending active learning with web technology*, Prentice Hall

Sadler, D.R. (1998) Formative assessment: revisiting the territory, *Assessment in Education*, 5(1), 77-84.

Winne, P. (2005), A perspective on state-of-the-art research on self-regulated learning, *Instructional Science*, 33, 559-565

Yorke, M. (1999), *Leaving early; undergraduate non-completion in higher education*. London: Falmer.

Yorke, M., and Longdon, B. (2004) *Retention and student success in higher education*. SRHE and Open University Press.

DEVELOPMENTS IN ON-SCREEN ASSESSMENT DESIGN FOR EXAMINATIONS

Che Osborne and John Winkley

Developments in On-Screen Assessment Design for Examinations

John Winkley and Che Osborne
BTL Group Limited
www.btl.com
che.osborne@btl.com

Abstract

This paper draws on examples from projects undertaken for a range of UK agencies, including the regulators from each of the 4 nations (QCA, ACCAC, SQA and CCEA), and Awarding Bodies such as Edexcel and the British Computer Society.

This work includes the use of:

- rich media (exploring how video, audio, animation and imaging affect assessment performance, including for candidates with disabilities),
- interactivity and adaptivity (exploring how requiring students to make interactive responses affects achievement and engagement),
- advanced computer-marking techniques (work to mark candidates' prose, mathematical workings, and process as well as output),
- item banking complex items to allow "when ready" assessment, and comparability issues with more traditional assessments.
- Working with authors across multiple locations and disciplines, and how the challenges can be met.

The paper also discusses how "when-ready" e-assessment is blurring the traditionally clear boundary between summative and formative assessment, and the opportunities open to qualification providers to reshape their assessment offerings to act as learning resources.

About BTL Group Ltd

BTL (www.btl.com) is a leading UK supplier of technology solutions for e-learning and e-assessment. In our e-learning developments, we provide a turnkey service for the design, scripting and production of learning packages, including components such as needs analysis, assessment, portfolio kits,

courseware and accreditation tools. In e-assessment we provide both the on-screen assessment content, and the delivery systems and services to Government Agencies and Awarding Bodies for use in both learning and examination settings This year we are launching our award-winning assessment content development system, CP3, which allows awarding bodies to develop and manage their own on-screen interactive assessment content.

Our UK customers for e-learning and e-assessment include DFES, DWP, QCA, BECTA, BBC, learndirect, RM plc, Edexcel and Pearson, OCR, the British Computer Society, SQA and the Teacher Training Agency.

BTL is independently owned and based in Saltaire (nr Leeds and Manchester), England. We employ approximately 75 staff. Our sister company, Virtual College (www.virtual-college.co.uk/), based in Ilkley, provides e-learning delivery services to industry in vocational and professional areas.

One of BTL's products described in this paper – CP3 recently won 2 awards at the British Computer Society Technology Awards. CP3's lead developer, Andrew McAnulla, won Young IT Practitioner of the Year Award, and the product itself was a medallist in the Best Products of 2005 - Service Products category.

The SQA Solar Project

SQA is an executive non-departmental public body sponsored by the Scottish Executive Education Department. It is the national body in Scotland responsible for the development, accreditation, assessment and certification of qualifications other than degrees. It is primarily funded through qualification entry charges and has an annual turnover of approximately £51m. It employs approximately 650 staff in Glasgow and Dalkeith and there are approximately 1,750 centres approved to offer our range of qualifications, including international centres.

The SOLAR Project (**S**cottish **O**n**L**ine **A**ssessment **R**esources) is funded by the Scottish Further Education Funding Council and is supporting the delivery of HN (Higher National) Qualifications. These qualifications consist of units which are traditionally assessed internally within colleges, followed by an external summative end-of-course assessment.

This is a well-established system and has many advantages, however marking pressures on tutors (who have to mark unit end assessments) coupled with consistency and quality issues with internally set and marked unit assessments discovered during post-hoc verification (which could then lead to unexpected results in the summative tests) meant that SQA

considered some possible improvements. We believe these improvements not only offer significant benefits to the community of learners and teachers involved, but they also illustrate the powerful beneficial effect that "next generation e-assessment systems" can have on Awarding Body relations with their customer centres, learners and tutors.

The project set out to provide a community-developed solution to the problem. Tutors in centres were invited to form "subject groups" with the strongest centres in each subject area taking the lead. These groups of tutors were then provided with technology and training which allowed them to develop on-screen objective unit assessments for the HN programme. These assessments are then submitted to SQA for Quality Assurance, before being signed off as live assessments. Centres (including those that authored the tests, and all the other Scottish FE colleges) then can provide these tests online to their candidature. The tests are electronically marked and results are available immediately. In addition, by pre-approving the tests, centres can offer them with confidence from the start of a course, with no risk of problems post-hoc with the validation.

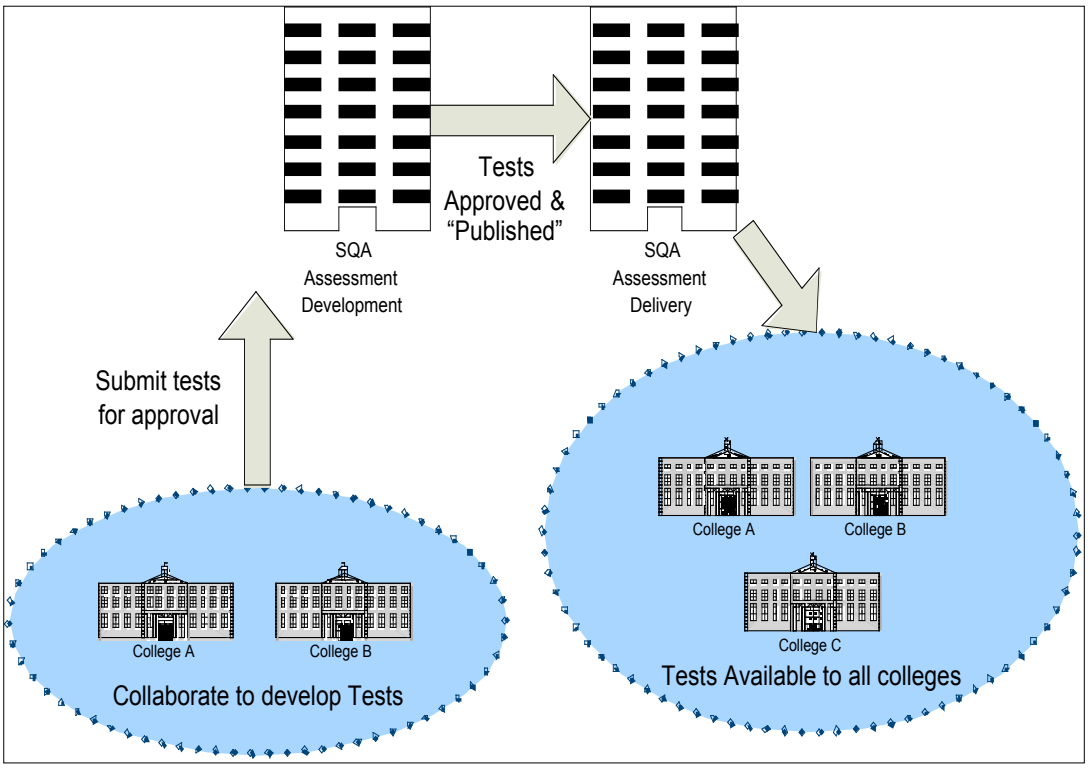


Figure 1 – Outline Process

Experiences in the Project

Broadly the project has been a success - it is now entering its 3rd phase, with approximately 50 colleges using 320 tests supplied by a community of 40 authors. By the end of the project we expect to have nearly 700 live tests on the system.

Throughout the programme, the implementation of the technology has caused considerably less problems than human factors – mainly communication and training. This is counter to what many expect to find – i.e. that the technology is now stable, but requires considerable skill in both using it, and applying it within the organisations. This has been particularly the case for the assessment development, where considerable training on both technical and educational (assessment design) aspects was required.

SQA and BTL's findings in the project are as follows:

- Training session on using CP3 authoring system and in assessment design is a constant and ongoing requirement – training at the outset is unlikely to be sufficient. The additional factor of multiple author communities in multiple locations, with multiple abilities brings multiple challenges.
- Customers and suppliers need a common understanding of project expectations and priorities.
- Success within the project was more about the suitability of the curriculum than technology (which broadly delivers as promised)
- There is no single eAssessment system that can provide all that a Qualification Authority requires
- Essential to adapt requirements based on user experience, and to work particularly hard on communication between all parties at all times. This has implications in terms of support and project management.
- The Invitation-To-Tender procurement process is problematic where the project concerned has evolving requirements (due to both lack of certainty at the outset and the inevitable experiences gained from running a highly innovative project).
- Having made these points, the experience of the authors concerned has been ultimately positive in that they believe they have learnt about e-Assessment, assessment design, and about their own subject area - an unexpected benefit of the project.

Supporting Innovative Assessment Delivery

As a supplier of exam systems recognised for their ability to support innovative assessments (both in terms of the content, and the delivery modes) BTL was interested in the SQA project because it offered the possibility of connecting development and deployment systems in a web-enabled setting.

In the UK, our experience is that first generation e-assessment projects generally start with replication of existing paper processes (this applies to both the test development and test delivery phases). In addition to the obvious

familiarity benefits of this (and therefore reduced risk in the technology requirements specification process) there are also advantages in terms of proving the comparability with paper tests, which often continue to run in parallel.

In subsequent phases, organisations begin to explore the specific benefits of on-screen assessments (in terms of efficiency and effectiveness gains). These are well documented in other projects but can include:

- Flexibility of delivery in terms of time, pace, place
- Immediate results: in addition to allowing rapid progression this can help bridge the traditional gap between formative and summative assessment: By providing tutors with immediate (and therefore useful) feedback about the detail of learners' performance in specific areas.
- Operational cost savings in centre.
- Supporting institutional objectives of leveraging use of ICT.
- Providing more valid assessments by assessing a broader range of skills/knowledge in more realistic settings.

In projects such as the SOLAR programme, although there are significant benefits from moving to on-screen development and delivery, the UK experience is that there is no desire to compromise on areas of assessment that have been seen as traditionally important. For example, the move from human-marked to objective computer-marked assessments is treated with careful scrutiny, and the introduction of computers brings an expectation among teachers and learners alike that the on-screen assessment will make good use of the interactive and rich media capabilities of modern computers.

BTL saw the critical technology requirements of 2nd generation assessment systems are as follows:

- Providing a **distributed test development process** that supports workflow among a community of people with different roles and skills.
- **Need to deploy development and delivery tools across an entire assessment enterprise** – becoming less of a project and more of a mainstream activity (although paper systems often continue in parallel, of course)
- Support for the key benefits of ICT in assessment:
  - **Rich media**. Self-evidently, computers can deliver a wider range of media types than paper. Most notable are the

following: animation and video (with play, pause, slow motion and replay), audio, and use of colour. Simple use can lead to significant validity improvement: e.g. much of the UK literacy curriculum is about observing and participating in face-to-face and telephone interaction with others. Paper is weak at conveying such scenarios with good face validity: the simple use of video and audio adds greatly to the validity.

- **Interactivity**. Interactivity is useful primarily in two ways. Firstly it allows candidates to give answers to more complex questions without necessarily having to write their responses down in text. Secondly it offers the opportunity for simulation systems. The ability of a learner to observe a system, manipulate some of its parameters, take further observations, draw hypotheses and test them out, etc. is a crucial feature of many curricula and is well-supported by on-screen interactive content.
- **Adaptivity**. As a subset of interactivity, the ability of a system to adapt to its users activities is of great interest in assessment. This can speed up assessment and also provide increased motivation for learners in formative settings.
- Advanced computer marking. Using advanced computer techniques to improve the range of assessments that can be marked electronically (for example, marking diagrams, free text, mathematical formulae and processes).

- **Powerful Item Banking** – to support the ongoing development of new items, and modification of existing items in a bank while the bank is also being used to generate live assessment content.
- **Supporting Formative Assessment alongside Summative Assessment**.

Alongside these benefits of on-screen assessment are significant, they bring potential problems which development and delivery systems must seek to deal with:

- Complexity, cost of development & trialling can increase
- Issues with accessibility for learners with disabilities may increase
- Technical deployment may become more challenging, for example raising the minimum system specifications for PCs, servers or network bandwidth.
- More learner and teacher preparation may be necessary to ensure students are aware of what they are expected to do, and how to operate the ICT properly in order to do it.

The outline structure of BTL's system is shown in the diagram below:

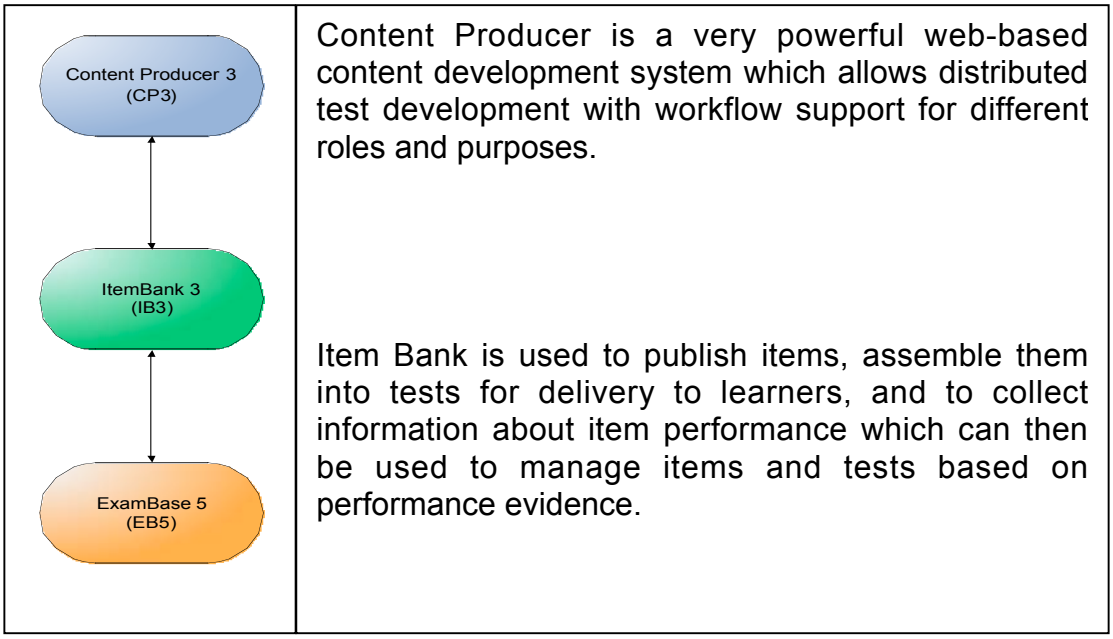


Figure 2 – Core Assessment System Components

The presentation which accompanies this paper will elaborate on some of the system's features and how they benefited SQA. The following features are particularly worthy of note:

- Item development ranges from the very simple to the very powerful. The development platform uses templates to allow rapid and simple creation of basic items, but leverages the full powers of Flash and XML to support more complex items, tests and curriculum taxonomies.

- The development phase is abstracted from the final delivery platform, allowing content to be produced and then published to a variety of output forms at a later date. This allows (for example) a bank of traditional items to be held in XML form and output to either on-screen or on-paper at the time of test assembly. It also allows practice tests to be published for delivery in other systems (e.g. within a VLE).

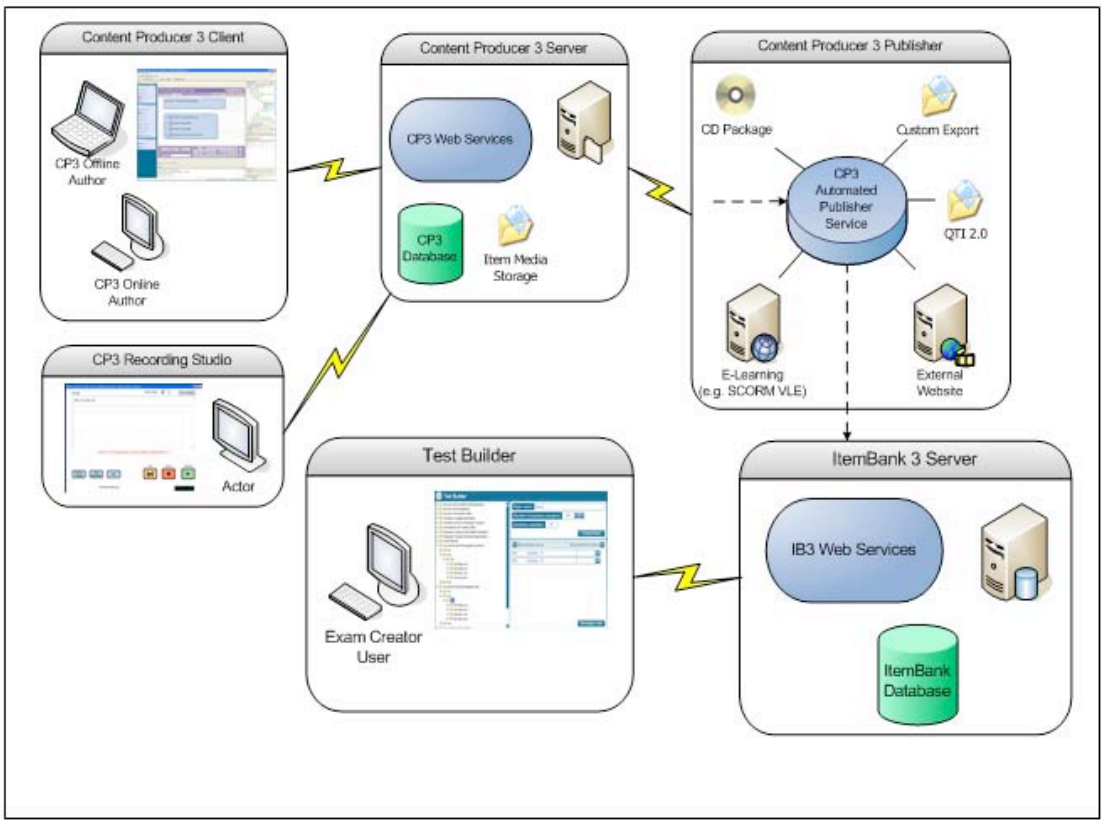


Figure 3 – System Functionality

- There are effectively two item banks. The first, part of the CP3 content production system is for items in development, at various stages in their workflow. These items are free to be edited according to the rules of the workflow and the user's role. Once published to the ItemBank IB3 Database, the item is fixed – potentially being used in live examinations and having candidate data stored about its performance. Modifications to the item must be made in the content development system and the 'new' item must then be republished.

- The rules for assembling tests (both static and dynamic, i.e. fixed form and containing randomised elements) are highly complex, and subject to user control. Considerable effort has been devoted to producing a user interface for this test construction process which is sufficiently powerful but simple enough to be used by a Subject Officer to manage an examination.

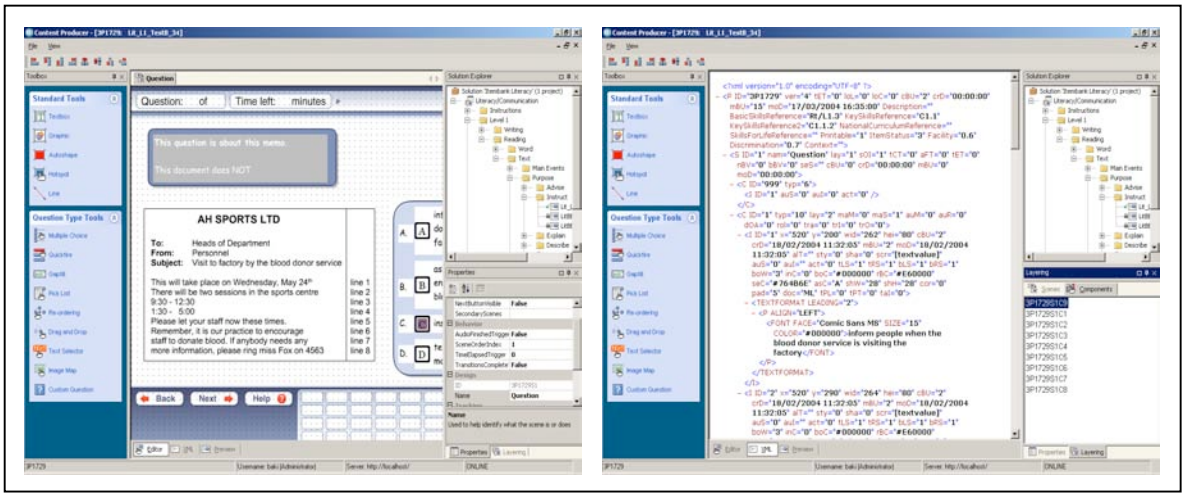


Figure 4 – Examples of CP3 Development screens showing XML and WYSIWIG Views

The CP3 development system is supported by a substantial team of developers and used by BTL's in-house production team for client content development (in fact the same system is used for e-learning and e-assessment content). However in deploying the system in customer centres (e.g. Awarding Bodies) to allow in-house content development, the additional supporting features have been required:

- A telephone and email helpdesk offering technical and assessment design support and advice.
- A maintained and supported FAQ and User Guides, including simple "How To" Tutorials for occasional users
- Template playbooks detailing all the (~150) item types that CP3 can support as standard.
- Systematic processes of qualifying trainees as capable to use the system. Currently we operate a 3 tier structure for CP3 producers with access to different features at each level, to ensure that users who are still learning do not stray into areas of "dangerous" functionality. This programme is supported by a series of tests and examinations (and these are used as part of the HR/personnel performance review programme within BTL).
- A carefully managed programme of upgrades. The CP3 system is under continuous development both to meet specific customer requirements (for example recent work includes improved support for accessible content and the ability to import and output QTI IMS v2.0 content). While upgrades for internal staff can be rolled out with informal communication, it is important that upgrades are both planned and notified in advance to avoid external users simply seeing additional or different features on the desktop.

Within the examination delivery system which accompanies CP3 (called ExamBase) we have seen rapid increase in both the volume of centres and the number of tests (the graph below demonstrates take-up on one of our customer's assessment programmes). Alongside this growth, we have seen a corresponding decrease in technical problems with installing new centres which we attribute to a combination of improved process and increasing user readiness for e-assessment.

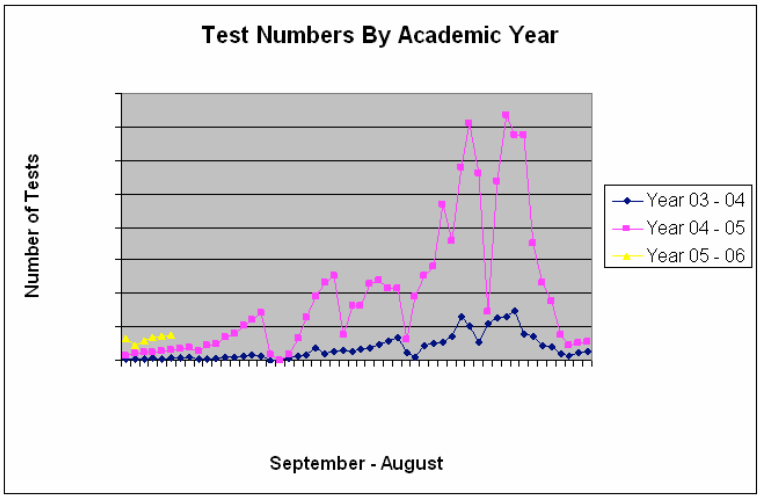


Figure 5 – E-Assessment Take-Up

Developments in Formative Assessment

Considerable work is underway (in parallel with e-examination development and deployment) to use the power of ICT to provide immediate powerful and detailed feedback from formative assessments which can be used as part of the learning process. One example of this is the suite of tools developed for the English Department for Education & Skills (Ministry of Education) for the Skills for Life Qualifications. Formative Assessments exist at each of the interventions in the diagram below.

Supporting the production of on-screen assessments by external authors where feedback frames are included is complex, as the feedback itself is effectively an additional set of conditional screens based on the marked outcomes of the questions. Our presentation will demonstrate recent examples of innovative work in this area.

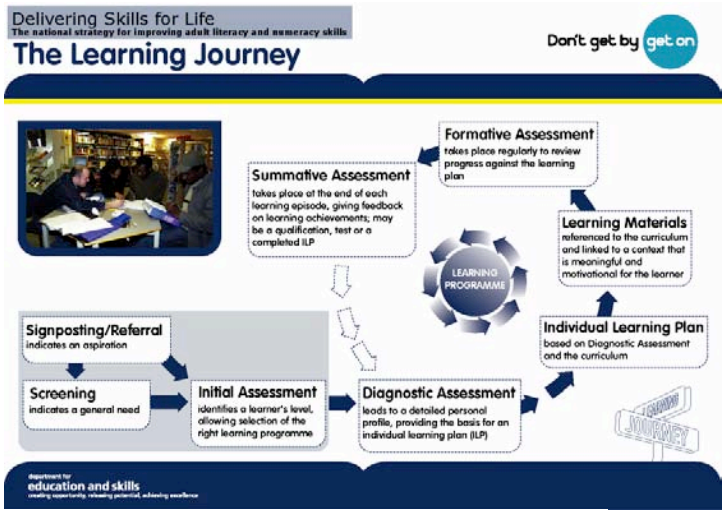


Figure 6 – The Learning Journey

One current view of how best to tackle formative assessment is set out by Black and Wiliam's "Working inside the black box" (Kings College, London), which holds out the promise of very significant achievement gains if the formative assessment techniques are used. However, the administrative burden of marking and managing large quantities of personalised assessment data is a real challenge for busy teachers.

Although quantitative marking is discouraged by Wiliam and Black (in favour of qualitative feedback), our experience with CAA is that candidates value immediate scoring (particularly for simpler, more objective assessments). In any event, computers are poor at qualitative feedback on longer pieces of work - essentially our findings are that in the absence of higher order formative assessment, which is difficult, immediate objective formative feedback, linked to a personal learning plan is both motivational and useful to learners.

There are a number of levels at which the feedback can take place:

1. It may refer to a group of questions, usually through a mark or a simple qualitative comment following some written responses.

2. It may refer to an individual question, following verbal questioning in a group or on an individual basis, either verbally or on paper.

3. It may refer to one step in a question, with the teacher looking over the shoulder of the learner and pointing out a mistake as it occurs, or marking a question with meticulous care.

All of the above take place in a traditional teaching and learning context, but limitations on teacher time mean that learners get more feedback at level 1 than at level 2, and in turn more at level 2 than at level 3. The opportunity

presented by e-learning is to provide much more feedback at level 3, because the computer does not have the limitations on time faced by the teacher.

In our view it is not realistic for the computer to provide feedback at level 3 of the traditional type ("explanation") except in very rudimentary form. This is because the number of possible responses required is vast (it is known as a combinatorial explosion), and cannot be programmed in. "Online Help" systems seem so wooden and stupid because of this problem.

On the other hand, it is much easier to track the learner's work electronically and highlight an error as soon as it occurs. This has the advantage of leaving the learner with the cognitive conflict, an important part of the learning process, and also a clear view of the precise location and nature of the problem. All this adds up to the ideal conditions for learning. Its nearest equivalent is a teacher looking over a learner's shoulder and pointing out a mistake as it occurs – but answering further questions with questions rather than explanations. The computer is ideally situated to deliver at least parts of this kind of Socratic Dialogue.

Our recent work in ICT-supported formative assessment seeks to provide the learner with immediate and relevant feedback at the point of error in order take advantage of both the elements of Wiliam and Black's recommendations regarding Assessment for Learning, and the lessons learned regarding the benefits of immediate results/feedback to learners in terms of achievement and motivation. In addition to helping the learner progress with a problem, advances in ICT-mediated Formative Assessment also hold out promise for classroom teaching - helping teachers to manage the large amount of performance information that the assessment is providing, thereby providing timely information to focus teaching effort.

We hope to present out initial findings from trials of these new assessments at the conference.

Future Developments

As the understanding of the impact of projects like those outlined above grows, the demands placed on systems, processes and suppliers continues to grow to meet every more sophisticated requirements. Leveraging technology without impacting on the core deliverables of a given project or diluting the assessments themselves becomes a key concern for organisations wishing to benefit from the adoption of industrialised e-assessment.

Whilst the above examples go some way to illustrating the ever more sophisticated demands being made of both technology and suppliers, there are additional areas worth noting as part of a vision for the future that do not deal strictly with technology.

Training

As e-assessment moves further towards the mainstream, there is a danger that the ability to leverage the full benefits that the technology and associated processes offer are over looked in the rush to handle the purely technology issues. Whilst many technology suppliers offer “point and click” based product training, it is felt that there is still a shortage of impartial pedagogy based e-assessment training. One of the key areas of growth will be the supply of material looking at areas such as:

- Writing onscreen questions (Impact of screen size, question types etc.)
- The importance and Impact of feedback
- The impact of transferring paper test's onscreen
- The importance of proper piloting to understand the above
- Statistics and their use for assessment compilation
- Adaptive test compilation, the benefits and challenges

Whilst this knowledge may be widespread at a conference such as this, it is BTL’s experience that this knowledge is not widely available or disseminated outside of those who might be classed as early adopters. Any organisation wishing to industrialise it’s delivery of onscreen test’s will need to address this knowledge gap, but may struggle to find the resources to do so.

In the coming year BTL will be working in conjunction with Alpha*plus* (www.alphaplusconsultancy.co.uk) to address this need, with pilot courses being run in September 2006. BTL would be keen to discuss this offering with any organisation that might wish to be involved or pilot this material.

Tendering

An additional area that continues to fail both suppliers and organisations adopting e-assessment is that of fixed price tendering. Over a short term small scale pilot project, the objectives for a given project might not alter significantly from those proposed at the outset. However, over longer term, higher stakes or more innovative projects, the ability to adapt to lessons learned during a project can significantly improve the likelihood of a successful outcome. The current position with fixed price tendering tends to mean that unless something was fully specified at the outset of a project, there is little scope to build in anything additional. An example of this might be that providing practice test’s might be seen to aid the learners ability to pass a final high stakes exam, but if this wasn’t specified or budgeted for from the outset, it might trigger another round of tendering for an organisation to be able to leverage this potentially important addition.

Whilst it is understood that the tendering process is in place to offer some certainty and protection to the purchasing organisation, it must also be recognised that this will place quite significant restrictions on how adaptive a supplier organisation can be. Although the widespread adoption of project management methodologies such as Prince2 have tools such as change control to combat some of these challenges, they so not offer a complete solution, as they rarely allow for budget movement outside of a pre-set tolerance.

One way of combating these challenges is to accept from the outset that expectations are going to change within the lifespan of a given project, and to allow for this. Some organisations have found it beneficial to move towards framework agreements with a list of preferred suppliers which can be used against a pre determined table of charges. This allows organisations to pre approve it's suppliers, understand how their charges are levied, and call those off as required. The freedom offered with this arrangement allows for organisations to expand or contract the scope of a project without having to re-tender for it's entirety, and also to potentially use separate suppliers for given pieces of a project on a mix and match basis.

ASSESSMENT TO IMPROVE SELF REGULATED LEARNING

Poppy Pickard

Assessment to Improve Self Regulated Learning

Poppy Pickard
Learning and Teaching Fellow
University of Bolton
Deane Road
Bolton
BL3 5AB
pp7@bolton.ac.uk

Abstract

This short paper considers how strategies of giving timely and enabling feedback, assist students in regulating their learning on a level 1 java programming module using a blended learning approach. The module has two short computer delivered assessments. Feedback for the programming exercises has been given 'face to face' instead of the previous method of VLE delivered feedback. The paper considers the effects of this change.

Introduction

At the University of Bolton many computing students study Java as their first programming language. The Java module has been running with relative success for 3 years (1, 2), students are given access to a variety of online materials, including animated learning objects, course notes, practical activities and so on.

The module has weekly assessed programming exercises, an end-of-module problem solving programming task and two short assessments during the module, each worth 20% of the coursework. Each short assessment which lasts for two hours and takes place during a practical session has two parts: a programming exercise and a multiple-choice quiz, both delivered through the VLE, WebCT. The multiple choice quiz selects questions randomly from a topic set, marks for the multiple choice quiz are released when the cohort has completed the quiz. Students can review the quiz in detail during the next practical session. Feedback and marks for the programming exercise are always available in WebCT by the following practical session.

Catalyst for Change

It has been noticed how in the past a few students have ceased to attend the module after these assessments. Over three years, with 500 students, on average 7.5% stopped attending after the first assessment and a further 6% after the second assessment.

This semester remediation is being attempted by piloting a different approach. Giving feedback comments in WebCT meant they were disassociated from the programming code and not always understood by the student. Feedback comments delivered in this way which are a transmission of the tutor's own view will most likely be first viewed by the student in a situation where the tutor is not present to share in a dialogue.

The New Approach

Writing a program under test conditions as a novice programmer can be a daunting experience. Before the test students were instructed to 'comment out' any lines of code they felt were incorrect rather than deleting them and leaving no evidence of their thought processes. In this way credit could be given for something that was partially correct. To improve the quality of learning through feedback, students were required to mark their own programs using a clearly defined solution and marking scheme which was e-mailed to all students once the task was completed by the whole cohort.

By using this methodology the intention was to adopt some of the seven principles of good feedback practice recommended by Nicol and Milligan (3).

1. helps clarify what good performance is (goals, criteria, expected standards);

2. facilitates the development of reflection and self-assessment in learning;

4. encourages teacher and peer dialogue around learning.

Principle 1. Giving students a solution and marking scheme, that rewards both good style as well as correctness, enabled the students to see the required standard for this assessment as well as understanding the marking process. This was particularly important in the second programming exercise where it was possible for a student to have a 'working solution' to the problem but one that was inefficient in programming terms.

Principle 2: Having the solution and being required to use it, required students to reflect and measure their own performance against a specified standard.

Principle 4: The process facilitated dialogue and understanding between the tutor and student.

The students were required to present their marked program the following week in the practical class, in order to receive their annotated and marked program from the tutor. These were then compared for similarity giving a basis for discussion where there was a significant discrepancy.

Discussions as Part of Feedback

The discussions enabled the student to see why their program was failing or how it could be improved. Programming is an activity that requires the programmer to pay attention to often minute details in the code. This attention to detail is well served by encouraging good habits in beginners, as often there are some novice programmers who simply want to 'make it work' and then move on to the next task. Some of these minutiae are about good style, i.e. adopting the appropriate conventions for the programming language, others are critical to the correctness of the program.

After the first programming assessment conversations centred more around issues of style, whereas after the second assessment dialogue focussed more on structural issues. In particular after the second assessment conversations highlighted how students needed varied feedback. Little feedback was needed for those who had already corrected their own errors in order to satisfy any frustration they felt in having a task that was incomplete. Others who were failing in the logical parts of the task needed the mediation of dialogue and gesture, i.e. pointing to and showing the amendments to the logical structures involved in order to be able to conceptualise their errors. Again using gesture and dialogue some needed to be shown a re-ordering to make their programs more efficient, it was not possible on the marking scheme to show how each inefficient order could be adapted.

Results

The programs were marked out of 20. After the first assessment, about 70% of the students marked within 2 marks of the tutor's mark, rising to about 80% after the second assignment. The prevalent trend for both assessments was for students to award less marks than the tutor.

Students were also required to complete a reflective questionnaire after each assessment.

Question	Test 1	Test 2
The mark reflected my programming ability	92%	88%
marking my own work helped me understand what was required	92%	88%
the tutor feedback was helpful	100%	100%
I was adequately prepared for the programming assessment	82%	88%
I felt confident whilst taking the test	90%	80%
Average mark for programming exercise (out of 20)	13.4	14.2

Percentages indicate those agreeing

The second programming assessment yielded broadly similar results to the first, except a about 6% felt better prepared and 10% felt less confident whilst taking the test. This was not however reflected in the average marks.

Module numbers and completions are given below. These have been recorded two weeks after second assessment in week 11. There are 67 students enrolled on the module of whom 57 have actively participated. The 10 excluded have either never attended or only attended once or twice at the beginning and not taken any assessments.

Test 1	Survey 1	Not seen after 1	Attended after 1, missed 2	Test 2	Survey 2	Missed 2, but attended since
54 (3)	39	2	3	43 (5)	26	4

Completions (bracketed numbers are students with mitigating circumstances)

Conclusions

Has the approach been successful?

There were 2 disappearances immediately after the first assessment. Comparing with previous figures this is 2 out of 55 (3.6%) and shows an improvement from the average 7.5% over the last three years. There is concern for the 4 students who missed the second assessment as yet for no given reason. On balance this is an improvement on previous semesters. Students are responding well to the detailed feedback and although this does not use a disproportionate amount of practical time, it does use more tutor time.

This approach has been used to replace on-line feedback; however there is a challenge to see if the feedback methodology can be implemented on-line and still maintain these improvements.

As the module is still live, there may be minor alterations in the data presented at the conference.

References

(1) Boyle T., Pickard P. et al, *Introducing a virtual learning environment and learning objects into higher education courses*, International Journal of Learning Technology, vol 1, nos4, 2005, pp. 383-398

(2) Boyle T., Pickard P. et al, *Using blended learning to improve student success rates in learning to program*. Journal of Educational Media, Special Issue on Blended Learning, vol 28, nos 2-3, October 2003, pp. 165 - 178.

(3) Nicol D. & Milligan C, (2006) *Rethinking technology-supported assessment practices in relation to the seven principles of good feedback practice*. In C. Bryan and K. Clegg (Eds), Innovative Assessment in Higher Education, Taylor and Francis Group Ltd, London (pub 22/03/06).

SYSTEM DYNAMICS BASED LEARNING ENVIRONMENTS: A TECHNOLOGY FOR DECISION SUPPORT AND ASSESSMENT

Hassan Qudrat-Ullah

System Dynamics Based Learning Environments: A Technology for Decision Support and Assessment

Hassan Qudrat-Ullah
School of Administrative Studies
York University
4700 Keele Street
Toronto
ON
Canada
M3J 1P3
hassanq@yorku.ca

Traditionally decision support systems (DSS) are designed to help the users make better decisions. However, the empirical evidence concerning the impact of DSS on improved decision making and leaning in dynamic tasks is equivocal at best. In this article, we introduce a new type of DSS based system dynamics technology as tool not only to support users' decision making and leaning but can also provide an effective assessment of the performance and learning as well.

Introduction

Managers face problems that are increasingly complex and dynamic. Decision support system (DSS) are designed to assist them make better decisions. However, the empirical evidence concerning the impact of DSS on improved decision making and learning in dynamic tasks is equivocal at best (Klabbers, 2003; Todd and Benbasat, 1999; Sharda et al., 1988; Sterman, 2000). Over four decades of dynamic decision making studies have resulted in a general conclusion on why people perform poorly in dynamic tasks. In dynamic tasks, where a number of decisions are required rather than a single decision, decisions are interdependent, and the decision making environment changes as a result of the decisions or autonomously or both (Edwards 1962), most often the poor performance is attributed to subjects' misperceptions of feedback. That is, people perform poorly because they ignore time delays between their 'actions and the consequences' (Sterman, 2000) and are insensitive to the feedback structure of the task system (Diehl and Sterman 1995). Decision maker's mental models about the task are often inadequate and flawed (Kerstholt and Raaijmakers, 1997; Romme, 2004). In this paper we argue that system dynamics based interactive learning environments (ILEs) could provide effective decision support for dynamic tasks by reducing the misperceptions of feedback. How do we know that learning has occurred? We argue that the design of ILEs facilitate the

automatic capture of decision making data and provides an effective learning assessment.

Background

Dynamic Decision Making

Dynamic decision-making situations differ from those traditionally studied in static decision theory in at least three ways: a number of decisions are required rather than a single decision, decisions are interdependent, and the environment changes, either as a result of decisions made or independently of them or both (Edwards, 1962). Recent research in system dynamics has characterized such tasks by feedback processes, time delays, and non-linearities in the relationships between decision task variables (Romme, 2004). Driving a car, managing a firm, and controlling money supply are all dynamic tasks (Diehl & Sterman, 1995) In these tasks, contrary to static tasks such as lottery type gambling, locating a park on a city map, and counting money, multiple and interactive decisions are made over several periods whereby these decisions change the environment, giving rise to new information and leading to new decisions (Forrester, 1961; Sterman, 2000).

ILE

We use “ILEs” as a term sufficiently general to include microworlds, management flight simulators, DSS, learning laboratories, and any other computer simulation-based environment – the domain of these terms is all forms of action whose general goal is the facilitation of dynamic decision making. Based the on-going work in the system dynamics discipline (Moxnes, 2004; Otto & Struben, 2004; Qudrat-Ullah, 2005b; Sterman, 2002), this conception of ILE embodies learning as the main purpose of an ILE. Under this definition of ILE, learning goals are made explicit to the decision-makers. A computer-simulation model is built to represent adequately the domain or issue under study with which the decision makers can experience and induce real world-like responses (Qudrat-Ullah, 2005a). Human intervention refers to active keying in of the decisions by the decision makers into the computer-simulation model via the interface of an ILE.

Performance in Dynamic Tasks

How well do people perform in dynamic tasks? The empirical evidence (Diehl & Sterman, 2000; Klabbers, 2003; Moxnes, 2004; Sterman, 2000) suggests almost a categorical answer: “very poorly”. Very often the poor performance in dynamic tasks is attributed to subjects’ misperceptions of feedback (Moxnes, 2004; Sterman, 2000). The misperception of feedback (MOF) perspective concludes that subjects perform poorly because they ignore time delays and are insensitive to feedback structure of the task system. The paramount question remains; are people inherently incapable of controlling system with time lags, non-linearities, and feedback loops? Contrary to Sterman’s MOF hypothesis, an objective scan of real world decisions would suggest that experts can deal efficiently with highly complex dynamic systems

in real life, such as, for example, manoeuvring a ship through restricted waterways. The expertise of river pilots, for example, seems to consists more of using specific knowledge (e.g., pile moorings, buoys, leading lines) they have acquired over time than in being able to predict accurately a ship's movements (Schraagen, 1994). This example suggests that people are not inherently incapable of better performance in dynamic tasks. Instead, decision makers need to acquire the requisite expertise.

Decision Making and Learning Assessment with ILEs

There exist some fundamental barriers to developing expertise in dynamic tasks: (1) *dynamic complexity*: our limited ability to understand the impact of time delays between our actions and their consequences coupled with the interactions between feedback loops that are multiple and non-linear in character and are ever present in the task systems we face in the real world, (2) *information availability limitations*: information we estimate, receive, and communicate is often oversimplified, distorted, delayed, biased, and ambiguous, (3) *information processing limitations*: when it comes to decision making people generally adopt an event-based, open-loop view of causality, ignore feedback processes, fail to appreciate time delays and are insensitive to nonlinearities present in the feedback loop structures of the task system, perceive flawed cognitive maps of the causal structure of the systems, make erroneous inferences even about the simplest possible feedback systems, fall prey to judgmental errors and biases, defensive routines and implementation failure (Sterman, 2000). The effective DSS, therefore, should allow the users to overcome such impediments to decision making and learning in dynamic tasks.

ILEs meet this challenge through the provisions of (1) a representative simulation model of the task system, (2) powerful interface, and (3) human tutor support--the three fundamental components of any ILE.

Decision Support through the Simulation Model

The greatest strength and appeal of an ILE in supporting decision making and learning in dynamic tasks lies in its underlying simulation model. In an ILE, the simulation model is built on system dynamics methodology (Forrester, 1961). The fundamental premise of system dynamics methodology is that 'the structure of the system drives its behaviour'. That structure consists of feedback loops, stocks and flows, and nonlinearities arising from the interaction of these basic structures (Sterman, 2000; Oliva, 2003). A typical system dynamics model allows that:

- The interaction and feedback between the systems variables, over time, in and across various sectors (e.g., demand, supply, production, finances etc.) of the task system be explicitly represented and the structural assumptions are made explicit and open.

- The disequilibrium framework for modeling be established, where the adjustments, say in the need for variable 'A' in response to the

changes in the variable ‘B’ to new equilibria typically crate imbalances and transient behavior.

- Delays and other distortions in perceiving the true value of the variables be explicitly modeled.

- Desired and actual variables magnitudes be explicitly distinguished from real magnitudes in the model.

- Non-linear responses to actions be explicitly represented.

The significance of the modelling capabilities of system dynamics methodology is its contribution to our understanding of the structure and behaviour of complex, dynamic systems. An understanding of the relationship between the structure (s) and behaviour (s) leads to the formulation of a better mental model of the task system (Sterman, 2002) and improved decision making (Brekke and Moxnes, 2003; Romme, 2004).

Decision Support through the Interface Design

Dörner (1980) asserts that decisions makers in dynamic tasks must acquire some reasonably precise notions of relationships among key task variables and develop an understanding of the most influential delays and feedback loops in the task system. System dynamics methodology provides powerful tools to represent qualitatively the connections between structure and behaviour of the task system through (i) causal loop diagrams and (ii) stock and flow structures. Utilizing these tools together with advances in modern IT, powerful interface, whereby references to the underlying simulation model are facilitated interactively, in an ILE can be constructed (for an excellent illustration please see, Romme (2004)). In this way, ILEs aid decision making by allowing the learners to examine the structure-behaviour relationship as and when needed in an ILE session.

Decision Support through Tutor Support

Decisional aid in the form of human tutor support constitutes the distinguishing and fundamental component of an ILE model. In an ILE session, decisional aids can be provided at three levels: pre-, in-, and post-task levels. Pre-task level decisional aids can be conceptualized as information provided by the human tutor to a decision maker about the model of the task prior to performing the task (Corner, Buchanan, & Henig, 2001; Davidsen & Spector, 1997). In-task decisional aids attempt to improve the individuals’ decision-making performance by (i) making the task goals explicit at early stages of learning, (ii) helping them keep track of goals during the task, and (ii) providing them with ‘diagnostic information’ (Cox, 1992). Post-task level decisional aids aim at improving performance by providing the decision-makers an opportunity to reflect on their experiences with task (Cox, 1992; Davidsen & Spector, 1997). Thus, an ILE could support the user’s understanding of dynamic tasks by offering the opportunity to, experimentally, design, test, and evaluate their decision strategies.

Learning Assessment with ILEs

In addition to their role as decision support and leaning tool, ILEs can be used as an evaluation tool as well. We have developed such an ILE, FishBankILE, in which learners have access to decision variables that determine their task performance and task knowledge. Subjects also have access to relevant information that may support their decision making and learning. The implementation of FishBankILE allows unobtrusive measurement of subjects' decisions and decision rules. For instance, FishBankILE's underlying simulation model automatically captures the task performance metric of the leaner using the following algorithm:

> *The task performance metric is chosen so as to assess how well each subject did relative to a benchmark rule (a built-in routine in FishBankILE system). The task performance measure for subject s, TP_s has the following formulation:*
>
> $$TP_S = \frac{\sum_{t=1}^{n_y} \sum_{t=1}^{n_T} \left| y_{it} - b_{it} \right|}{n_y * n_T}$$
>
> *where n_y is the number of performance variables, n_T is the number of trials the task has to be managed, b_{it} is the benchmark value of performance variable i at time t, and y_{it} is the empirical value of task performance variable i at time t. Task performance, TP, is assessed in the following way. Every decision period, the benchmark's performance variables' values are subtracted from the subject's. The subject's final performance, TP, is the accumulation over 30 periods of this difference, averaged over the number of task performance variables and number of trials*

In the next step of our project, we intend to use FishBankILE to asses the learning of students as well as professional program participants at our school.

Conclusion

Dynamic decision making research is highly relevant to both in-class learning and the managerial practice (Diehl & Sterman, 1995; Kerstholt & Raaijmakers, 1997). We need effective DSS to help the managers cope with the ever-present dynamic tasks. We presented ILE as a viable decision support and learning evaluation tool. Investigations regarding the overall effectiveness of ILEs, we believe, will advance our insights into the design conditions for an effective DSS to promote decision support and learning assessment in a variety of context.

References

Brekke K. A. & Moxnes, A. (2003). Do numerical simulation and optimization results improve management? Experimental evidence. Journal of Economic Behavior and Organization, 50(1), 117-131.

Corner, J., Buchanan, J., & Henig, M. (2001). Dynamic decision problem structuring. Journal of Multicriteri Decision Analysis, 10(3), 129-143.

Cox, R. J. (1992). Exploratory learning from computer-based systems. In S. Dijkstra, H. P. M. Krammer, & J. J. G. van Merrienboer (Eds.) Instructional models in computer-based learning environments (pp. 405-419). Berlin, Heidelberg: Springer-Verlag.

Davidsen, P. I. & Spector, J. M. (1997). Cognitive complexity in system dynamics based learning environments. International system dynamics conference. Istanbul, Turkey: Bogacizi University Printing Office.

Diehl, E. & Sterman, J. D. (1995). Effects of feedback complexity on dynamic decision making. Organizational Behavior and Human Decision Processes, 62(2), 198-215.

Dörner, D. (1980). On the difficulties people have in dealing with complexity. Simulations and Games, 11, 8-106.

Edwards, W. (1962). Dynamic decision theory and probabilistic information processing. Human Factors, 4, 59-73.

Forrester, J. W. (1961). Industrial dynamics. Cambridge, MA: Productivity Press.

Kerstholt, J. H. & Raaijmakers, J. G. W. (1997). Decision making in dynamic task environments. In R. Ranyard, R. W. Crozier & O. Svenson (Eds.) Decision making: Cognitive models and explanations (pp. 205-217). New York, NY: Routledge.

Klabbers, J. H. G. (2003). Gaming and simulation: Principles of a science of design. Simulation & Gaming, 34 (4), 569-591.

Moxnes, E. (2004). Misperceptions of basic dynamics: The vase of renewable resource management. System Dynamics Review, 20, 139-162.

Oliva, R. (2003). Model calibration as a testing strategy for system dynamics models. European Journal of Operational Research, 51, 552-568.

Otto, P. and Struben, J. (2004). Gloucester Fishery: Insights from a group modeling intervention. System Dynamics Review, 20(4), 287-312.

Qudrat-Ullah, H. (2005a). MDESRAP: a model for understanding the dynamics of electricity supply, resources, and pollution. International Journal of Global Energy Issues, 23(1), 1-14.

Qudrat-Ullah, H. (2005b). Behavior Validity of a Simulation Model for Sustainable Development. International Journal of Management and Decision Making, (forthcoming).

Romme, A. G. (2004). Perceptions of the value of microworld simulation: Research note. Simulation & Gaming, 35, 427-436.

Schraagen, J. M. C. (1994). What information do river pilots use?, Report TNO TM 1994 C-10, Soesterberg: Human Factor Research Institute.

Sharda, R., Steve, H., Barr, J., &McDonnell, C. (1988). Decision support system effectiveness. Management Science, 34(2), 139-159.

Sterman, J. D. (2000). Business Dynamics. New York: McGraw-Hill.

Sterman, J. D. (2002). All models are wrong: Reflections on becoming a system scientist. System Dynamics Review, 18(4), 501-531.

Todd, P. & Benbasat, I. (1999). Information Systems Research, 10, 356-381.

MATHEMATICAL QUESTION SPACES

Christopher J Sangwin

Mathematical Question Spaces

Christopher J Sangwin
Maths Stats and OR Network
School of Mathematics
University of Birmingham
Birmingham
B152TT
C.J.Sangwin@bham.ac.uk

Introduction

It is uncontroversial to assert that learning mathematics is only effective when it is an active process on the part of the learner. Setting questions is a ubiquitous technique to engage students, and answering such questions constitutes a large proportion of the activity they undertake. Indeed, asking students questions is a central part of all theories of learning.

This paper examines in detail the process of randomly generating versions of mathematical questions for CAA. In doing this we examine not only a single mathematical question, but how such questions are linked together into coherent structured schemes. Two important pragmatic reasons are often cited by colleagues for wishing to generate a random sequence of questions.

- Randomly generated questions may reduce plagiarism
- Distinct but equivalent questions may be used for practice

Even if giving each student a distinct problem sequence reduces plagiarism, professional experience unfortunately demonstrates it is not eliminated. However, some students are well aware of the potential benefits of collaborative learning, possibilities for which are traditionally hard to provide in the mathematics classroom. As one student commented in their feedback evaluations:

"The questions are of the same style and want the same things but they are subtly different which means you can talk to a friend about a certain question but they cannot do it for you. You have to work it all out for yourself which is good."

Notice here the student voices the opinion that the questions *"want the same things but they are subtly different".* In this paper we address exactly this issue, by examining equivalent mathematical problems in some detail.

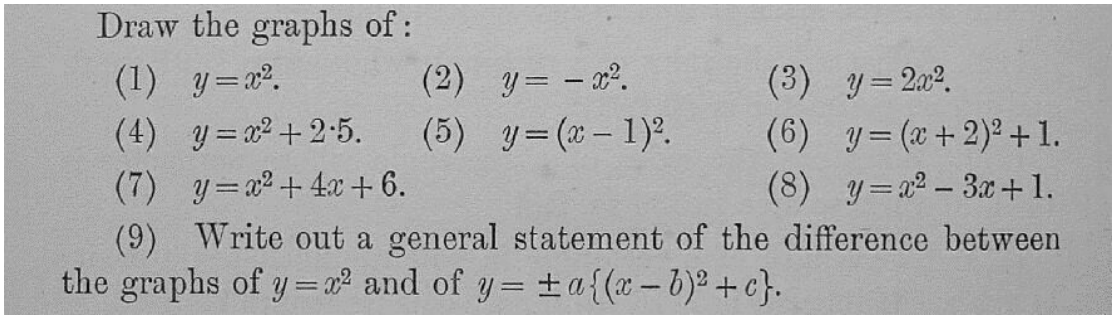
Mathematical Questions

Linguistically, a question is a sentence worded or expressed so as to elicit information. We shall use the term "question" in such a way, when in practice many words are used in text books, for example “exercise", "problem", "task" and even “examples". Here, a question is also taken to include an *instruction*, such as “solve", "factor", “sketch" and so on.

Using schemes of questions is one of the major techniques used for self-study, home work or in the classroom. Working through such pre-structured exercises is akin to taking part in a dialogue, and such dialogues are an important part of learning. Although it is usual for a dialogue to take place between two interlocutors, an internal conversation occurs when one engages in "thinking aloud". On the nature of this internal conversation [7] says, *"the mere act of communicating our ideas seems to help clarify them, for, in so doing, we have to attach them to words (or other symbols), which makes them more conscious"*. Hence, while one does not have a conversation *with* the textbook, the textbook may provoke internal enquiry and dialogue. They may also play a part in the learning process by providing mutual ground, or shared sequences of experiences, about which subsequent conversations can take place. There may be other legitimate uses, such as providing "finger exercises" to promote rather mindless, but nevertheless important, mechanical fluency.

A crucial distinction, when considering a mathematical question, is whether or not one cares about the answer. With many questions, no one cares about the actual answer. The purpose of the question is either to (i) practise some technique, or (ii) help build or reinforce some concept by prompting reflective activity. In other cases the purpose of the question is to obtain the answer. The question itself is a prototype of a practical problem which may be encountered, and hence this result may be useful.

We begin our examination of mathematical questions with a sequence of simple questions from [9]. This small, unassuming volume consists of 178 pages. There is no text or worked examples, instead simply sequences of problems. *"These examples are intended to provide a complete course of elementary algebra for classes in which the bookwork is supplied by the teacher"*. Part of one such sequence is shown in Figure 1.



Draw the graphs of:

(1) $y = x^2$. (2) $y = -x^2$. (3) $y = 2x^2$.

(4) $y = x^2 + 2 \cdot 5$. (5) $y = (x-1)^2$. (6) $y = (x+2)^2 + 1$.

(7) $y = x^2 + 4x + 6$. (8) $y = x^2 - 3x + 1$.

(9) Write out a general statement of the difference between the graphs of $y = x^2$ and of $y = \pm a\{(x-b)^2 + c\}$.

Figure 1: [9], pg 62

We claim that this sequence of questions is highly structured, and this example has been included here because clues to this structure are revealed in the unusual final synoptic question. Further that the purpose of such a sequence of questions is to develop concepts rather than obtain an answer or practice technique. Note however that including such a question may make little sense for a student who has struggled with questions (1)-(8), and has little work of merit from which to form a coherent synopsis. Question 1 provides a base from which comparisons can be made. Questions 2 and 3 re-scale the y-axis. Question 4 is a vertical shift, question 5 is a horizontal one, and question 6 involves both. Questions 7 and 8 also require simple shifts, although some simple algebra is required to reveal precisely what these are.

Many books contain word problems where part of the process is setting up the equations themselves. This is modelling, in its broadest sense. Some of these problems are practical, others mathematical. What they have in common, is that the answer appears to be applied, and hence it is the answer which is important. They do not appear to be conceptual, nor for practice. Rather they might be termed utilitarian. The following (admittedly somewhat dated) example is taken from [1]. However similar (if not identical) examples may be found in many modern books.

Examples XXVII. b. 10. If 6 fewer bottles of wine can be bought for £5 when the price is raised ten shillings per dozen, what is the original price?

In many cases such exercises are highly structured, with examples carefully chosen to reveal different cases in the underlying mathematics.

A third category of questions are those which seek to practice some skill. For example, [1] Chapter XIV contains some 325 repetitive exercises on the topic of factoring quadratics alone. This large quantity of repetitive practice is typical of many algebra books, including modern ones. These sequences of problems tend to be highly structured. This structure includes things which are common to whole sequences of problems, for example integer roots, the signs of the roots are all positive, and things which are varied.

As a concrete example of constrained variation consider the following question.

$$\text{Solve } ax^2+bx+c=0. \qquad (1)$$

We might consider indexing the individual instances by using coordinates *(a,b,c)*. Clearly, there are some subspaces, such as the subspaces of mathematically possible questions. The subspace satisfying b2 3 4ac characterizes the question subspace with real solutions. While such a mechanical indexing of questions is technically feasible, we would like to consider a quite different issue. This is to draw an analogy with the concept of an *example space* developed by [10]. An example space is taken to be the cognitive domain possessed by the student, rather than some intrinsic mathematical space. We seek to develop a dual notion: that of *mathematical question space*. Just as with example spaces, the notions of the *dimensions*

of possible variation and *ranges of permissible change* in any question space appear to be very useful. Each dimension of possible variation corresponds to an aspect of the question which can be varied to generate a collection different question instances. The range of permissible change is more problematic. "Permissible" may of course be taken to indicate the strict mathematical criteria of well-posedness, or may be used in a pedagogic sense. Given our educational context, a question space is considered to be the collection of instances which are educationally equivalent. That is to say, two instances in a space differ in ways which do not alter the purpose or effect of a question within that particular scheme. Furthermore, we identify the mathematical question with this pedagogic question space. While the student is likely to be aware only of the task in hand: the question instance, to the teacher this instance actually represents the question space and hence the underlying generality.

Clearly, the question space is more complex than simply varying a coefficient in a term. For example, in question 7 of the problem set shown in Figure 1, the question is an instance of a quadratic with no real roots, for which the completed square form is tractable. An instance of such a question would probably be given as an expansion of $(x-a)^2+b$, where a is a small integer, and $b > 0$ is a small integer. Hence, a particular dimension of variation certainly does not correspond to the direct variation of a coefficient in a question instance. As a result, to implement randomly generated instances from a question space sophisticated tools are necessary.

Clearly here it is easy to identify how the dimensions of variation affect the question instances, but it is unlikely that such an algebraic clarity will be evident in many situations. Equally, there is nothing to suppose that a dimension of variation will be algebraic at all. Variation could include which variable is used, the dimensions and orientation of geometric shapes, or the adjectives used in a word problem. Furthermore, there are many situations when a parameter will remain within a question, perhaps to suggest to the student that there is a range of permissible and “essentially the same" examples encapsulated within one question. It is possible in some circumstances that a question space will only contain one instance. For example, in Figure 1, question (1), there may be no reasonable alternatives, and the question space consists only of the instance "Draw the graph of x^2".

While practice of some technique could be seen to be the repeated completion of question instances from a particular question space, we argue that it is not. A selection of questions usually shows progression through a sequence of slightly different cases. Each of these will be consciously different, and so will be instances from different question spaces.

Existing Standards for CAA

In this section we consider the data model for the representation of questions for CAA provided by the IMS Question & Test Interoperability (QTI) specification. For them, an item is the smallest self contained exchangeable assessment object.

“An item is more than a 'Question' in that it contains the question and instructions to be presented, the response processing to be applied to the candidates response(s) and the Feedback that may be presented (including hints and solutions)."

Such a concept of a self contained item is present in virtually all CAA systems, either at an explicit or implicit level. In their sense it is significantly more than a question, since it contains details of response processing instructions, and feedback, both hints and solutions, to be given. This specification includes the notion of Item Clone, which are equivalent items created from an Item Template by the substitution of Item Variables. However, the specification operates only at the level of individual items, and takes no account of the sequence of items.

Similarly, the IMS Simple Sequencing Specification provides a mechanism for representing the intended behaviour of a "learning experience", the prototype of which is interactions with a sequence of items.

We argue that for mathematics the split between “item" and “sequence" is artificial and fails to capture crucially important aspects of the learning process in automated assessments built upon it. While it will be necessary to author and store items at this level, there is no clear distinction at the pedagogic level between item and sequence and it is often actually difficult to decide what the smallest exchangeable object is. Is a multi-part item a collection of separate items? While mathematics assessment can be shoe-horned into this data representation model, the results are unsatisfactory.

The STACK CAA System

This section concerns the implementation of a computer aided assessment (CAA) system for mathematics known as STACK: a System for Teaching and Assessment using a Computer algebra Kernel. A demonstration server is available at (http://www.stack.bham.ac.uk). As the names implies, STACK relies on a computer algebra system (CAS) at its heart to support a variety of tasks. The most important feature is that the CAA system evaluates the student answers containing mathematical content, rather than allow selection from a list of teacher provided answers, such as in multiple choice or multiple response questions.

Systems under which the processing of student answers is supported by computer algebra have gradually gained ground in higher education over the last five years. Perhaps the first system to make CAS a central feature was

the AiM system, described by [2], with subsequent technical developments described in [8]. This system operates using Maple, as does the Wallis system of [3]. Other systems have access to a different CAS, such as CalMath which uses Mathematica, CABLE, see [4], which uses Axiom and the STACK system which uses the CAS Maxima. From private correspondence, the authors are also aware of systems which use Derive in a similar way.

Details of the question authoring process are given in [6], and the important issue of student input syntax in [5]. From our point of view we are most interested in random generation of structured mathematics questions. Experience with STACK and similar CAA systems demonstrated that virtually all necessary tasks can be performed with the following three functions, when backed up by the sophisticated library of CAS functions.

- Generate a random integer between 0 and n.
- Generate a random floating point number between 0 and n.
- Select a random item from a list.

The important issue is the availability of CAS, or CAS-like, functions which can be used to build structured mathematical objects. Describing this at a level of detail suitable for interpretability is a difficult task, and one unlikely to be completed in the near future.

STACK, as with the vast majority of contemporary CAA systems, currently only operates at the level of individual items. While it is clear how richer multi-part items can be developed, it is not clear how technically separate but pedagogically connected items can be linked, to aid exchange and efficient re-use. This is the subject of ongoing work.

References

[1] H. S. Hall. A School Algebra. MacMillian, London, 11th printing edition, 1929. First published 1912.

[2] S. Klai, T. Kolokolnikov, and N. Van den Bergh. Using Maple and the web to grade mathematics tests. In Proceedings of the International Workshop on Advanced Learning Technologies, Palmerston North, New Zealand, 4-6 December, 2000.

[3] M. Mavrikis and A. Maciocia. Wallis: a web-based ILE for science and engineering students studying mathematics. In Workshop of Advanced Technology for Mathematics Education in the 11th International Conference on Artificial Intelligence in Education, pages 505-512, Sydney, Australia, 2003.

[4] L. Naismith and C. J. Sangwin. Computer algebra based assessment of mathematics online. In Proceedings of the 8th CAA Conference 2004, 6th and 7th July, The University of Loughborough, UK, 2004.

[5] P. Ramsden and C. J. Sangwin. A liberalised mathematical syntax for computer-aided assessment. In Proceedings of the International Mathematica Symposium, Perth, Australia, 2005.

[6] C. J. Sangwin and M. J. Grove. STACK: addressing the needs of the "neglected learners". In Proceedings of the WebAlt Conference, Eindhoven, 2006.

[7] R. R. Skemp. The psychology of learning mathematics. Penguin, 1971.

[8] N. Strickland. Alice interactive mathematics. MSOR Connections, 2(1):27-30, 2002. http://ltsn.mathstore.ac.uk/newsletter/feb2002/pdf/aim.pdf (viewed December 2002).

[9] C. O. Tuckey. Examples in Algebra. Bell & Sons, London, 1904.

[10] A. Watson and J. Mason. Extending example space as a learning/teaching strategy in mathematics. In A. D. Cockburn and E. Nardi, editors, Proceedings of the Annual Conference of the International Group for the Psychology of Mathematics Education (PME26, Norwich, United Kingdom), volume 4, pages 378-385, 2002.

MOODLE: ENHANCING THE ASSESSMENT CAPABILITIES OF THE LEADING OPEN SOURCE VIRTUAL LEARNING ENVIRONMENT

Niall Sclater, Phil Butcher, Pete Thomas, Sally Jordan

Moodle: Enhancing the Assessment Capabilities of the Leading Open Source Virtual Learning Environment

Niall Sclater n.l.sclater@open.ac.uk
Phil Butcher p.g.butcher@open.ac.uk
Pete Thomas p.g.thomas@open.ac.uk
Sally Jordan s.e.jordan@open.ac.uk

The Open University

With the merger of Blackboard and WebCT, the selection of the open source system *Moodle* is an increasingly attractive alternative for many institutions. The Open University (OU) recently choose Moodle as a core component of its virtual learning environment after an extensive requirements gathering process and evaluations of commercial and open source products. The University has now launched a £4m programme to enhance the Moodle suite of e-learning tools, integrate Moodle with existing systems and promote the uptake of the new tools by course teams. It is feeding back its developments to the Moodle community and in turn hopes to reap the benefits of continual efforts taking place across the World to enhance the pedagogical provisions of the system.

A key Moodle module being enhanced by the OU is the Quiz Engine. While it has some good features e.g. ease of question authoring with an immediate preview facility, ability to define a range for numeric variables in numeric questions, and randomised questions in a test, it currently has a limited range of question types, does not fit well with University quality assurance and exception handling processes and is weak on feedback. Enabling better feedback is a particular concern as the OU has always paid attention to the role of assessment in the learning process. Our own in-house assessment system, OpenMark, has been designed to support the provision of detailed personalised feedback and to allow multiple attempts at each question thereby enabling students to receive feedback and act on it immediately. An initial assessment has been made of the potential for using Moodle to provide these more complex question types and we have concluded that it is possible to include such questions, and their feedback, within Moodle tests. Work is now being carried out to determine whether, or how, some of the other features of OpenMark, such as feedback on competences (evidenced from answers for a group of questions) can be built into Moodle.

The overall conclusion is that, while the Moodle quiz module does not currently meet OU functional requirements, it is proving feasible to substantially enhance and integrate it with other OU systems. Indeed, a common concern faced by institutions with well-developed but diverse systems is the effort required to incorporate them into a full VLE. Therefore, the OU is keen to pursue this interfacing in conjunction with the worldwide Moodle community.

This paper argues that there are many advantages in using an assessment system which fully integrates with an institutional virtual learning environment. It reports on the requirements gathering that has taken place, outlines the development work currently under way, examines some of the challenges and advantages of developing software as part of a global open source community and proposes future changes to the way assessments are handled within Moodle which should be of interest to the wider Moodle community. It also reports on the issues the OU is examining surrounding interoperability, accessibility, handling of maths questions, automated text marking, adaptive testing and item banks.

INNOVATIONS IN E-ASSESSMENT

Eric Shepherd

Innovations in e-Assessment

Eric Shepherd (CEO)
Questionmark Computing
info@questionmark.co.uk
+44(0)8007315895

The cornerstone of successful education is the effective use of assessments. The 21st century offers a real opportunity to use technology to make assessments more widely available and more successful for those involved in the process. In a world where you cannot know everything, assessments will be used to guide people to powerful learning experiences, reduce learning curves, confirm skills, knowledge and attitudes, and motivate by providing a sense of achievement.

Since launching its first computerised testing product nearly two decades ago, Questionmark has been at the forefront of e-assessment technology. Join Questionmark CEO Eric Shepherd to learn about user-driven innovations in e-assessment and how they will benefit education professionals.

The Questionmark™ Perception™ assessment management system enables educators to create questions and organise them into exams, quizzes, tests or surveys. Administrators can schedule students to take the assessments, deliver them in a variety of ways and then view the results in multiple different report types. Role-based security and workflow management enables multiple authors work collaboratively.

In 2005 Questionmark introduced exciting new authoring, security and content management capabilities. Over the past year, Questionmark has introduced dozens of new features, capabilities, and integrations to meet the assessment management needs of thousands of education and assessment professionals worldwide. The newest version of the Perception combines a new reporting system and administrative features with many other enhancements that make it easier to manage large numbers of participants, administer assessments at certified test centers and provide a better overall participant experience.

This session will explain and demonstrate some of the new technologies that will be help education and assessment professionals author, deliver, monitor, and report on an increasing number of assessments easily and securely including: new reporting capabilities for assembling, formatting, saving and distributing meaningful reports from your assessment data; item searching

that enables authors to quickly find, update and manage large item banks; participant “browser checks” to ensure reliable delivery, allowing a participant to log in to an assessment only when a compatible browser and configuration are detected; test center management tools make it possible to schedule tests for specific test centers and to require proctor log-ins for high stakes tests; intuitive user interfaces for managing participants and schedules make it easier to find and create schedules for groups, participants, assessments; and administrative functions that provide more control and flexibility for scheduling tests and accommodating participants with special needs. The session will also explore new developments that allow integration of assessment management with best-of-breed course management and open source systems. Finally, the session will introduce you to a new program that enables assessment professionals to share and exchange item banks.

Join us for an informative and interactive session on how the latest innovations in e-assessment authoring, management, delivery and reporting can dramatically enhance the way educators can use assessments to measure knowledge, skills and attitudes.

EVALUATING THE USER EXPERIENCE IN CAA ENVIRONMENTS: WHAT AFFECTS USER SATISFACTION?

Gavin Sim, Janet C Read & Phil Holifield

Evaluating the User Experience in CAA Environments: What Affects User satisfaction?

Gavin Sim and Janet C Read
Department of Computing
University of Central Lancashire
Preston
PR1 2HE
Tel: 01772 895162
grsim@uclan.ac.uk
jcread@uclan.ac.uk

Phil Holifield
Faculty of Design and Technology
University of Central Lancashire
Preston
PR1 2HE
pholifield@uclan.ac.uk

Abstract

This paper reports the findings of an experiment to establish students' satisfaction with various aspects of the user interface in three Computer Assisted Assessment (CAA) environments. Forty four second year undergraduate students in Human Computer Interaction participated in the study. Each student completed three tests using three different CAA software environments. Through the use of two survey instruments, user satisfaction was measured. The results highlight the fact that, in this instance, scrolling did not seem to influence student satisfaction but other attributes, such as navigational structure and question styles, appear to influence it. The students appeared to prefer different CAA environments depending on whether the context of use was for formative or summative assessment.

Introduction

With the increased adoption of CAA within educational institutions there has been a rise in the number of such systems available. Several of these are designed for use in Higher Education establishments; these include Questionmark Perception, Hot Potatoes, TRIADS and TOIA. These 'bespoke' systems are relatively new to higher education but software delivering multiple choice style questions dates back to the 1970's (Morgan, 1979) and so the concept is quite old. In addition, there are learning management systems, like WebCT and Blackboard, that have CAA tools incorporated into them.

In a commercial marketplace it is important for the vendors of CAA software to attract new customers and then to hold onto their customer base. To attract new custom, vendors often emphasise the 'features' of their products, placing great importance on the number of different question styles available. In common with many other software products, with each new version, more features and more question styles are offered. For example, TRIADS software developed by Derby University offered 17 question styles in 1999 (Mackenzie, 1999) compared to 41 in 2005 (CIAD, 2005). It has been reported that instructors and academics are often unfamiliar with many of these highly sophisticated new question styles and subsequently find it difficult to write questions that take advantage of their features (McLaughlin, Fowell, Dangerfield, Newton, & Perry, 2004).

One user group that has little influence on the design of CAA software is the student population that uses the software for assessment. This group is seldom in a position to choose which CAA software is used and yet their experience of the software is clearly important. User experience is one of the facets of usability which is generally measured by considering the effectiveness of an interface, the efficiency of the system and the user experience (ISO, 1998). It is expected that the user experience of the software would have some impact on the test performance (Bridgeman, Lennon, & Jackenthal, 2002), however, there has been very little research analysing the user experience of CAA and in particular the effect on user experience when more sophisticated questions are introduced into the test environment.

The user experience is often related to the user satisfaction of a system and is concerned with how well the system facilitates the user in achieving their goal. User experience can be ascertained by the use of surveys and observations, that rely to some extent on opinions and judgements, as well as more scientific methods, these include measures of skin sweat rate and heart rate. The most common method for evaluating user experience is, however, the written questionnaire (Johnson, Zhang, Tang, Johnson, & Turley, 2004; Van Veenendaal, 1998).

Using questionnaires to gather user opinions is problematic, studies point to the tendency of individuals to choose random answers, to report what the questioner wanted to hear, and to fail to complete questionnaires (Vaillancourt, 1973). Careful design of questionnaires can reduce these issues, paying attention to the length of the survey as well as the length of the questions and adding questions that test for reliability are known solutions (Breakwell, Hammond, & Fife-Schaw, 2000).

In this study, questionnaires were used as the means to elicit opinions from undergraduate students about the user interface for three CAA applications.

Method

An experiment was devised using three CAA applications that provided between them a variety of interface design characteristics which the users could evaluate. At the outset of the experiment, there were several hypotheses about the impact of certain 'features' of CAA software with respect to user satisfaction. User satisfaction was considered to be affected by the user experience of:

- Accessing and finishing the test
- Navigation within the test
- Visual layout
- Interface for answering questions

In a CAA environment, the goal of the user is to complete the assessment, progression towards this goal requires the completion of several tasks; to start the test, answer the questions, navigate between pages and end the test (Sim, Horton, & Strong, 2004). It was expected that there would be some variation between CAA applications with respect to the above constructs. The purpose was not to identify, or claim, that one application was better than another, merely to examine attributes of the interface that affect user satisfaction. This limitation was necessary as the three applications being considered could be customised to present the tests in different formats and the students were examining the interaction within the environment and so were not using the software to test their knowledge of a specific subject domain.

Choice of CAA Applications

As outlined in the introduction, there are numerous CAA applications. For this study a choice was made to focus on three software applications, S1, S2, and S3. S1 was selected as an example of a CAA application integrated into a Learning Management System (LMS) as an assessment tool. Such tools usually have limited question styles compared to more specialist CAA software, however they are widely used for assessment purposes within Higher Education (Alexander, Bevis, & Vidakovic, 2003; Cooper, 2002; Pretorius, 2004; Sayers & Hagan, 2003).

S2 is a dedicated CAA software application offering a lot more functionality and question styles than learning management systems. Many institutions have adopted such software for formative and summative assessment (Sim, Holifield, & Brown, 2004).

Finally S3 is a CAA software application offering more advanced question styles than the other two applications and is perceived to be more flexible and specialist. A demonstration version of S3 was used exhibiting a variety of sophisticated question styles.

Software Set Up

For S1 (see Figure 1), the test was set up so that all the questions were displayed on the screen at once and three question styles were used; Multiple Choice, Multiple Response and Text Entry.

For S2 the test was set up using question by question delivery and incorporated the following question styles; Multiple Choice, Multiple Response, Order, Text Entry, Matrix and Drag and Drop.

Finally within S3 four sections of the demonstration were selected to be used which incorporated a variety of sophisticated question styles such as drawing lines, assertion reason and matrix.

Figure 1: Screen shots of the three software used (from left to right S1, S2, S3)

Survey Design

The study used two survey tools. The first (Q1) was a questionnaire adapted from an earlier version (Sim & Holifield, 2004) which had previously been used to examine user satisfaction with the interface of a CAA software application. Additional questions were included in Q1 to examine the effectiveness of the software in facilitating the user in achieving their goal.

This questionnaire (Q1) consisted of 13 Likert style questions and was divided into four sub-sections. To minimize acquiescence, the tendency by some of a sample to consistently agree or disagree with a set of questions (Bryman, 2004), a mixture of positive and negative statements were incorporated into the design. There was also the opportunity for students to provide qualitative data with regards to specific features they liked about the interface.

The second survey instrument (Q2) was a variation on a repertory grid (Fransella & Bannister, 1977) and loosely based on an instrument that is used for children to measure fun (Read, MacFarlane, & Casey, 2002). This was presented to the students one week after completing the evaluations of the three applications and it required the participants to rank each application according to nine constructs. This survey also included two questions that required the students to identify which of the CAA applications would be their preferences for formative and summative assessment.

Apparatus

The students conducted the first part of the experiment in three different labs using networked PCs with flat screen monitors, full size keyboards and scrolling mice. In each lab, the hardware specification was the same.

Participants

The students that took part in the study were a convenience sample taken from an undergraduate class in HCI. A total of 44 participated in the experiment, but only 25 completed the second survey (Q2). This class comprised students from seven different computing courses and therefore had a wide range of different 'types' of student for example, networking and software engineers. The sample was predominantly male and approximately 5% of the sample did not have English as their first language. The participants did not receive any payment for taking part in the study but a draw was made at the end of the experiment and the lucky winner got a free text book. Participation was voluntary but some may have felt it was a part of their class as it took place in class time.

Procedure

The evaluation of the CAA applications took place on a single day at a single time in three identically equipped computer labs. In these labs, students worked through a series of questions in the 3 applications. The order in which they met the three packages was counterbalanced to remove any learning effects that might otherwise have affected the results. Thus, in one lab everyone started with S1, in another everyone started with S2 and in the third, everyone started with S3. The S1 application had 17 questions on football, S2 had 17 questions on Films and S3 used the default questions from the online test interface which included topics such as geology and maths. Students worked through the three applications in their own time (but were supervised). They were able to move through the three applications in their own time but had to stick to the pre determined order. As each student completed a single application, they completed the questionnaire Q1.

For the post hoc study, students were given the repertory grid activity Q2 and asked to complete it. This was done in a class a week after the initial experiment. It was not possible to link these results to the results from the experiments.

Analysis

The first questionnaire, Q1, completed after the test was scored in an ordinal way 1-5, where 5 represented Strongly Agree and 1 Strongly Disagree. If the question was negatively worded then the scoring was reversed.

The Repertory Grid (Q2), completed the week after the initial experiment, was again coded in an ordinal manner using 1-3 for each of the criteria. The last two questions on the sheet “Which of the three would you choose for: an end of year exam” and “Which of the three would you choose for: Revision purposes” were tallied according to how many students selected that software.

Friedman tests were conducted to establish whether there were any significant differences between the three software applications and Wilcoxon post-hoc tests were then preformed to determine where the difference lay.

Results and Discussion

As the results reported in this paper are predominantly gleaned from the survey instruments, a test of reliability was carried out on the major instrument, Q1; the alpha reliability of the scale is 0.888.

In Q1, the students were asked whether they had any prior experience of using the software. From the 44 participants, 17 had prior experience of S1, 20 had experience of using S2, and only 2 had used S3 before. A Mann-Whitney U Test was conducted between those who had prior experience and those without for S1 and S2. There was no significant difference between the two groups on any of the questions, therefore prior experience does not seem to influence there satisfaction of a CAA environment.

The mean scores relating to the participants answers for Q1 are displayed in table 1. Overall on the majority of questions they reported a level of satisfaction with each of the three CAA environments.

No	Question	S1	S2	S3
1	I had no problem gaining access to the test	4.21	3.84	3.53
2	I encountered difficulties starting the test	4.28	3.95	3.40
3	The interface required too much scrolling	3.44	4.19	3.81
4	The amount of scrolling was acceptable	3.37	3.95	3.60
5	It was difficult to read the text on the screen	3.86	3.84	3.09
6	The screen layout was clear	3.88	3.67	2.56
7	The screen layout was consistent	4.12	4.02	2.72
8	I liked the way the test looked	3.49	3.33	2.35
9	I would have preferred an alternative font	3.40	3.23	2.86
10	The button names are meaningful	4.02	3.88	3.33
11	I always knew where I was within the software	4.02	4.05	2.23
12	The navigation was logical	4.05	3.84	2.65
13	The navigation was clear	3.95	3.77	2.58

Table 1: The mean scores for the first questionnaire for each of the three software applications

The results from the REP grid (Q2) which was administered a week after Q1 are displayed in table 2 below.

Student Ranking	S1	S2	S3
Login	18	6	1
Navigation	12	11	2
Layout	6	16	3
Scrolling	6	8	11
Reading	12	11	2
Instructions	9	14	1
Input Answer	11	11	3
Change Answer	11	9	0
Finish Test	12	12	1

Table 2: Frequency each piece of software was ranked first by the user on a number of criteria

Accessing and Finishing the Test

The first two questions in Q1 refer to the students gaining access and starting the test. Although the mean scores suggest overall there was little difficulty in accomplishing this task S3 was significantly different to S1 for the first question ($Z=-2.882$, $p>0.01$) and S3 was significantly different to both S1 ($Z=-3.987$, $p>0.001$) and S2 ($Z=-2.293$, $p>0.05$) for the second question.

Similar results were obtained in Q2 with S3 appearing quite different from the other two as only one student ranked it first. In addition, in this survey a post hoc Wilcoxon revealed a significant difference between S1 and S2 ($Z=-2.562$, $p<0.01$). The high scores for S1 could have been due to the fact that the majority of students access the LMS for teaching material for their modules and so the look, if not necessarily the test environment, was familiar to them. These differences may have also been as a consequence of the amount of interaction that is required before the user gets to the first question: S1 and S2 both required 5 tasks whilst S3 required 6.

Using Q2 the students were asked about how easy it was to end the test and only one student ranked S3 the easiest whilst S1 and S2 were both ranked easiest by 12 students. This may be because of the amount of interaction for exiting the test was higher in S3 than the other two applications.

Visual Layout

Both S1 and S3 incorporated scrolling in the user interface, in S1 the questions were all displayed on the screen and for S3 the scrolling was in the instructions and results. In Q1 there were two questions that examined the effects scrolling had on user satisfaction. For question 3, S2 was significantly different to both S1 ($Z=-3.473$, $p<0.01$) and S3 ($Z=-2.215$, $p<0.05$) and a similar result was obtained for question 4. However, Q2 revealed no significant difference between the three software in relation to scrolling.

The three applications all used different font types and sizes and this was presumed to affect legibility. In the first survey question 5 asked about the legibility of the text and in the answers to this, S3 was found to be significantly different to both S1 ($Z=-3.007$, $p<0.01$) and S2 ($Z=-3.15$, $p<0.01$) and similar results occurred for question 9. Q2 also asked about legibility and it was also found that S3 was scored lower than S1 and S2. This perception about the legibility of the text within S3 may have been because, due to this application being evaluated with the ready made questions rather than the simple style questions used in S1 and S2, there was a lot more text in both the questions and the feedback than in S1 and S2. This, coupled with scrolling which is a known factor that effects on screen legibility, could have led to the poor result for S3 (Bernard, Chaparro, Mills, & Halcomb, 2003).

Questions 6, 7 and 8 in Q1 also related to the layout of the screen and again satisfaction with S3 was significantly lower than S1 and S2. For example question 6, a Wilcoxon test revealed that S3 was significantly different to S1 ($Z=-4.463$, $p<0.001$) and S2 ($Z=-4.337$, $p<0.001$) with similar results found for questions 7 and 8. These findings were all supported by the results from Q2 where S3 was ranked lower than both S1 and S2. This may have been attributed to the fact that each question in S3 used a different style and therefore there was no continuity in the interface compared to the other applications.

Navigation

The final four questions in Q1 related to the navigation of the CAA software applications and again there were differences between them. For question 10 S3 was significantly lower than S1 ($Z=-3.018$, $p<0.01$) and S2 ($Z=-2.485$, $p<0.05$) this was also found to be the case for the other three questions relating to navigation.

The results from Q2 in relation to navigation revealed a significant difference $\chi 2=21.68$, $p<0.001$ and post hoc tests revealed that S3 was ranked significantly lower than S1 ($Z=-3.273$, $p<0.01$) and S2 ($z=-3.855$). There was no difference between the navigation of S1 and S2. The low results for S3 may have been due to the linear navigational structure, students being required to select an option then work through the questions in order. There was little freedom to move between questions or skip a question and return to it later.

Answering the Questions

Q2 asked the students about inputting an answer and there was again a significant difference between the three applications $\chi^2=19.76$, $p<0.001$. The post hoc test revealed there was no difference between S1 and S2 however, S3 was significantly lower than S1 ($Z=-3.855$, $p<0.001$) and S2 ($Z=-2.805$, $p<0.05$). These results were similar to the results relating to instructions and it is possible that because the level of interaction was more complex, students found the process of answering questions more difficult within S3.

Preference for Software Depending on Context

The final two questions in the survey asked the students which software they would choose for summative and formative assessment, the results are shown in Table 3.

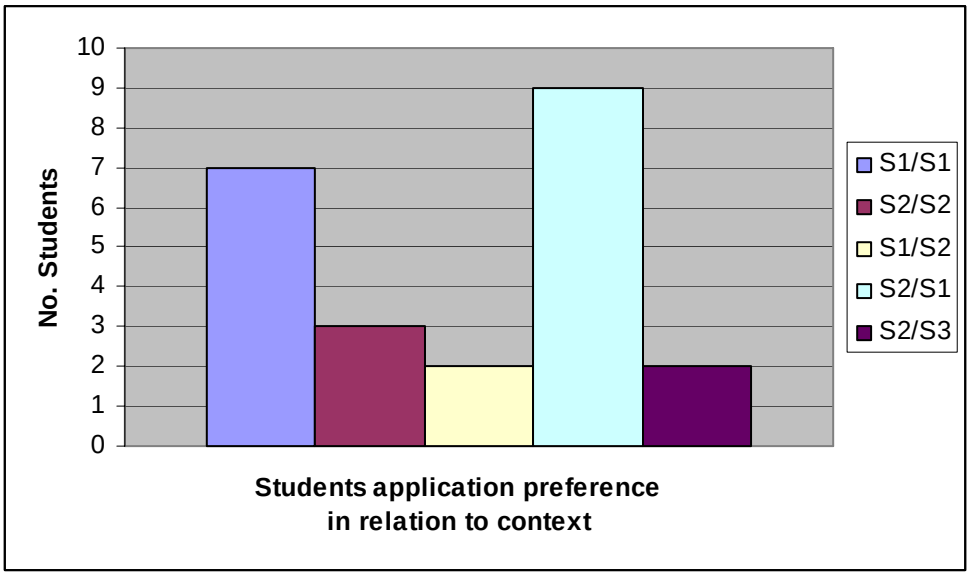


Table 3: Students application preference in relation to context

Of the 23 students completing this section only 10 stated they would use the same application for both contexts. The remaining 13 had a different preference depending on the context of the assessment. For example, 9 students stated their preference for summative assessment would be S2 and S1 for formative assessment. This would suggest the nature of assessment also influences students' perception about the suitability of a CAA environment and it is not just simply looking at the interface attributes.

Conclusions and Further Work

For developers of CAA environments or academics customising templates, this research has highlighted a number of interface characteristics that affect user satisfaction within a CAA environment. For S1 and S2 prior experience had no bearing on user satisfaction, it was not possible to examine this for S3 due to the limited number of students who had prior experience. It may be that

for more complex interaction prior experience is necessary to improve overall satisfaction as there is a greater learning curve.

There does not appear to be a single attribute that influences students' preference for a particular CAA environment. Other research has highlighted scrolling as an attribute that affects students attitude (Ricketts & Wilks, 2002) but in this study S1 required the most scrolling yet students indicated that they would still select this system for formative or summative assessment. With regards to navigation, students appeared to prefer the ability to navigate freely and were less satisfied with the linear structure presented in S3.

Increasing the number of question styles did not seem to affect attitudes between S1 and S2, however the complexity of the questions within S3 may have affected the students satisfaction. Further work may be needed to determine whether there is a complexity threshold within CAA environments in relation to question styles, and if so, whether once this threshold is passed, there is a related decline in overall user satisfaction.

When selecting and evaluating a CAA environment, context appears to be a significant factor that needs to be considered. Students appear to prefer different systems depending on whether the software is being used for formative or summative assessment. In this study, the majority of students selected S1 for formative assessment but this may be because they associate this application (which is part of a LMS) with their learning, considering S2 and S3, both more specialised and more suitable for assessment. There was a mixed response in relation to summative assessment with students opting to use either S1 or S2.

This study has highlighted the complexity of trying to do a comparative study of three CAA environments. In this instance it was not possible to customise S3 as a demo version was used; this undoubtedly influenced the results as apportioned to individual applications and so the results presented here cannot be used to indicate a preference or otherwise for a particular application. The intention of the study was to examine the interactions within general CAA environments.

There are several extensions to this work, it would be useful to ask students why they chose a particular application for formative and summative assessment, to determine what features they consider to be the most necessary and to investigate the effects of multiple question styles.

References

Alexander, M., Bevis, J., & Vidakovic, D. (2003). *Developing Assessment Items using WebCT.* Paper presented at the World Conference on E-Learning in Corporations, Government, Health and Higher Education, Phoenix.

Bernard, M. L., Chaparro, B. S., Mills, M. M., & Halcomb, C. G. (2003). Comparing the effects of text size and format on the readability of computer-displayed Times New Roman and Arial text. *International Journal of Human-Computer Studies, 59*(6), 823-835.

Breakwell, G. L., Hammond, S., & Fife-Schaw, C. (2000). *Research methods in psychology* (second ed.): Sage.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2002). *Effects of Screen Size, Screen Resolution and Display rate on Computer-Based Test Performance.* Paper presented at the Annual meeting of the national council on measurement in education, New Orleans.

Bryman, A. (2004). *Social Research Methods* (Second Edition ed.). Oxford: Oxford University Press.

CIAD. (2005). *TRIADS Question Styles*. Retrieved 22/11/05, 2003, from http://www.derby.ac.uk/ciad/triadstyles.html

Cooper, C. (2002). *Online Assessment using Blackboard an issue paper*: University of Wales Institute Cardiff.

Fransella, F., & Bannister, D. (1977). *A manual for repertory grid technique*. London: Academic Press.

ISO. (1998). *Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on usability*: ISO 9241-11.

Johnson, T., R., Zhang, J., Tang, Z., Johnson, C., & Turley, J., P. (2004). Assessing informatics students satisfaction with a web based courseware system. *International Journal of Medical Informatics, 73*(2), 181-187.

Mackenzie, D. (1999). *Recent Developments in the Tripartite Interactive Assessment Delivery System (TRIADS)*. Retrieved 13/06/02, 2002, from http://www.derby.ac.uk/ciad/lough99pr.html

McLaughlin, P. J., Fowell, S. L., Dangerfield, P. H., Newton, D. J., & Perry, S. E. (2004). Development of computerised assessments (TRIADS) in an undergraduate medical school. In D. O'Hare & D. Mackenzie (Eds.), *Advances in computer aided assessment* (pp. 25-32). Birmingham: SEDA.

Morgan, M. R. J. (1979). MCQ: An interactive computer program for multiple-choice self testing. *Biochemical Education, 7*(3), 67-69.

Pretorius, G. (2004). *Objective testing in an E-Learning Environment: a Comparison between two systems.* Paper presented at the World Conference on Educational Multimedia, Hypermedia and Telecommunications, Lugano.

Read, J. C., MacFarlane, S. J., & Casey, C. (2002). *Endurability, Engagement and Expectations: Measuring Children's Fun.* Paper presented at the Interaction Design and Children, Eindhoven, The Netherlands.

Ricketts, C., & Wilks, S. J. (2002). Improving Student Performance Through Computer-Based Assessment: insights from recent research. *Assessment & Evaluation in Higher Education, 27*(5), 475-479.

Sayers, H. M., & Hagan, N. S. J. (2003). Supporting and Assessing First Year Programming: The use of WebCT. *Italics, 3*(1), 1-11.

Sim, G., & Holifield, P. (2004). *Piloting CAA: All aboard.* Paper presented at the 8th International Computer Assisted Assessment Conference, Loughborough.

Sim, G., Holifield, P., & Brown, M. (2004). Implementation of computer assisted assessment: lessons from the literature. *ALT-J, 12*(3), 215-229.

Sim, G., Horton, M., & Strong, S. (2004). *Interfaces for online assessment: friend or foe?* Paper presented at the 7th HCI Educators Workshop, Preston.

Vaillancourt, P. M. (1973). Stability of children's survey responses. *Public opinion quarterly, 37*, 373-387.

Van Veenendaal, E. (1998). *Questionnaire based usability testing.* Paper presented at the European Software Quality Week, Brussels.

A CALL TO ARMS FOR HANDHELD DEVICES

Jon Trinder, Jane Magill and Scott Roy

A Call to Arms for Handheld Devices

Jon Trinder and Jane Magill
Robert Clark Centre for Technological Education
Scott Roy
Department of Electronics and Electrical Engineering
University of Glasgow
Glasgow
G12 8QQ
J.Trinder@elec.gla.ac.uk

Introduction

What are the obstacles preventing the widespread use of mobile devices for CAA? This paper is intended to stimulate discussion on how to best introduce mobile devices to learners, and how to provide optimal support for mobile CAA. The discussion is an opportunity to more clearly delineate obstacles and therefore arrive at better solutions to the challenges of mobile CAA in FE and HE. Some of these are common to the introduction of CAA in general, others subtly different because of the nature of the delivery platform.

Background

Mobile learning utilises devices such as Personal Digital Assistants (PDAs), Smartphones, and Media players to deliver educational material and facilitate learning. Delivery of CAA on these platforms is unique due to device characteristics: the small screen size; varying device form-factors; input mechanisms, with some having touch screens and others miniature keyboards – and the non-traditional learning environments where devices are used. The potential of mobile devices in education is widely recognised: *"…mobile devices can become efficient and effective teaching and learning tools" (Roibas and Sanchez, 2002). It has been predicted that "In future, learners need not be tied to particular locations. They will be able to study at home, at work or in a local library or shopping center, as well as in colleges and universities"* (Sharples, 2000a).

In the primary and secondary education sector a 2003 report from BECTA and DfES into the use of handheld computers in schools noted *"Handheld Computers (PDAs) could bring important benefits to schools by assisting administration, supporting classroom management and enabling personal and group learning"* (Perry, 2003). In schools, the provision of computers is not as good as in HE and pupils often have to access a computer at scheduled times. The BECTA report notes *"A Further benefit of the small size of PDAs is that they can be accommodated in any classroom on a one-each basis"*.
This lack of availability of personal desktop machines may therefore make PDAs attractive to pupils, with the resultant familiarity engendering other uses

of the devices, and an exploration of their full potential. It is therefore unsurprising that the most publicly visible large scale projects have, so far, been in schools such as those in Wolverhampton, where up to 1000 pupils will be given PDAs (www.expresso.co.uk, 2005).
The 2005 JISC Landscape Report into the Use of Wireless and Mobile Technologies in Post-16 Education notes *"Student experience with mobile devices in schools is likely to have an impact on their expectations for similar use in post-16 education"* (Evans, 2005). FE and HE need to be as innovative as the schools in utilising new learning technologies.

Indeed there are many potential benefits for students in higher education if mobile learning can be successfully deployed. With many students working part time the use of mobile devices – with their obvious portability and 'instantly on' functionality – gives opportunities for learning at non-traditional times. This has been confirmed for specific learning situations. *"Our team carried out a detailed study of how radiology is taught and practised. …computer-based learning must fit into the gaps in their busy schedule - in the hospital, at home, when travelling - which means a personal and portable system."* (Sharples, 2000b). The proliferation of WiFi capabilities in PDAs, and increasing ubiquity of WiFi provision in coffee shops, service stations, pubs, etc. opens the way for network based CAA resources to be utilised. Our research has also shown that PDA use actually enhances small group activity amongst learners, as participants work face-to-face rather than facing a computer screen. We also recognise that an additional advantage of PDAs is their relative affordability, allowing devices to be loaned to students in class or lab situations.
The future for mobile CAA seems promising, with some favourable outcomes (Attewell, 2005), and an active mobile learning community in the UK, mainly focussed on F.E and H.E., with forums and where ideas and research are discussed by a number of small research teams, (www.jiscmail.ac.uk/pda-edu www.handheldlearning.co.uk). However, after a number of years trialling PDA use in undergraduate settings, we have met with mixed success, and it appears that this is not unusual. One of the problems is that to prove that mobile devices can work in FE and HE requires they are used enough, and for long enough to realise their potential, but this means convincing both users and stakeholders that mobile learning *can* work. *"Undoubtedly there is a threshold to cross which requires sufficient immersion in any new technology to reach a point where it is of unquestionable value"* (Perry, 2003).

Barriers to Adoption

It would appear that mobile learning is an ideal tool for use by FE and HE students, and initial barriers to adoption would initially appear to be predominantly technical – how to translate pre-existing content for mobile CAA use, and how to transfer content to learners.

There are already a number question banks making use of the IMS/QTI format, but discussions with mobile application developers indicate that many find the format too complex and so only support considerably simpler formats in their software. General, automated, question re-authoring from such complex formats is non-trivial, and so re-authoring may either need to be done manually at considerable expense, or questions from other, simpler sources (eg a local VLE) used.

Furthermore, if students are not to be granted or loaned a pre-setup PDA, the diversity of the mobile devices they may already own makes it difficult to provide material that can be used on all of them. This problem might be addressed by limiting device support to the more functional devices (PDAs under PalmOS and Windows Mobile) and developing simple meta-formats easily translated for specific software on each device. The properties of such meta-formats, and the means of keeping a consistent GUI over the device range are still points of argument. An even simpler method might be to present material in the simplest common denominator format of simple text files – or other proprietary formats which already have cross platform compatibility – and accept less rich content and a far degraded ease of use for the learner.

In addition, as the mobile device market is continually reinventing itself (with considerable advertising push), the range of devices that may be supported increases. For example, should the rise of the ubiquitous music playback on PDAs and mobile ‘phones indicate that questions and answers may be better provided to students as podcasts? Does CAA have to be provided in the latest technologies and formats in order to be embraced by learners?

To transfer data to mobile devices there are various methods depending on the type of device: ‘beaming by infrared or Bluetooth, ‘syncing’ via a cable to a desktop machine, supplied on removable media (e.g. SDCard), transferred by WiFi if enabled, or sent as an SMS message.

Many of these will require either extra hardware and/or software to be installed in labs. Altering of computing labs is something which normally requires co-operation and approval from the departmental/institutional computing support officers and often takes a long time to be agreed to and implemented. In the future Wifi would seem a good choice, but many institutions are still reluctant to allow ad-hoc connection to their networks and many have not got used to connecting fairly standard student laptops so a diversity of mobile devices may not be welcomed or supported for some time.

In our pilot studies at Glasgow we have found that even when many of the technical barriers have been overcome inter-personal, personal and social factors can have a major impact. For instance, we have found that a factor in the slow acceptance of mobile devices is how well provided students are with open access labs. For campus based students all the information they need is readily available without the need for the reward/effort tradeoffs of working with a small screen device. In a recent study we conducted some students chose not to use the mobile device as it was easier to wait until they got home to use a normal PC. This may not be the case for students based away from campus or part time students integrating their study with full time employment.
Other social and personal factors we have noted include: the disruption of pre-existing group hierarchies on introducing PDAs, with existing group leaders experiencing a 'loss of face' due to a lack of proficiency with the new technology, and more junior students hiding their proficiency; students unwilling to accept technology perceived to be out of date compared with the technology they already owned (grey scale PDAs, compared with colour screen mobile phones with objectively lower computing power). The principles are summarised in the JISC Landscape document, *"lack of success may be due to inappropriate use for a given context, loaned devices may lose the benefits of personalisation, and students may abandon their use of mobile technologies if they believe their social networks are under attack"* (Kukulska-Hulme *et al.*, 2005).

A Killer Solution?

In the electronics industry the concept of the 'killer application' is prevalent: the piece of software so obviously or addictively useful that it of itself persuades users to purchase and use a new piece of technology. The relatively short history of mobile CAA, has seen a number of technologies posited as the 'killer solutions' which will enthuse learners. Presently the iPod is a 'must have' device for many students and as its functionality is extended it may become a useful mobile learning device (in addition to the present limited use in downloading podcasts). Devices such as the Sony PSP have potential for use as a mobile learning tool, having a built in browser, WiFi, and imminently, a Macromedia Flash player.
Yet it may be that the very personal nature of devices, one of the strengths of the mobile electronics industry, may mean that there will not be one 'must have' application for everyone, but that multiple possibilities for mobile CAA may have to be provided to suit individual, and institutional circumstances.
It seems likely in future that institutions will have to cater for a wide variety of computing devices: laptops, tablet PCs, ultra-portable PCs, iPods and other devices. As wireless access points become more common, students will expect to be able to access learning resources from wherever they are. Many of the problems mentioned in this paper such as the production of suitable content for different platforms, enabling connection to their network of various devices, will have to be addressed in order for institutions to attract students and remain competitive.
We hope that this paper will form the basis of discussion, and whether that discussion leads to solutions that do form a 'killer app' for mobile CAA, or, as

is more likely, ideas on how best to adopt mobile CAA into one's own specific institution, it may have instrumental in moving the field forward.

Points for discussion

How best to translate existing content for mobile CAA?
(Content changes to aid translation and usability, supported device subset)

How best to transfer content to learners?
(wired/wireless, central/distributed dissemination, simple/multimedia, intermittently/continually online)

Social barriers.
(Which social barriers have we encountered, and which are critical)

Staff & Institutional perception of mobile CAA.
("those little screens are too small to do anything with", etc.)

References

Attewell, J. (2005)(Mobile technologies and learning- A technology update and m-learning project summary http://www.lsda.org.uk/cims/order.aspx?code=041923&src=XOWEB (Accessed 24/02 06)

Evans, D. (2005)(Potential Uses of Wireless and Mobile Learning http://www.jisc.ac.uk/uploaded_documents/Potential%20Uses%20FINAL%202005.doc (Accessed 18/02 06)

Kukulska-Hulme, A., Evans, D. and Traxler, J. (2005)(Landscape Study in Wireless and Mobile Learning in the post-16 sector - Summary http://www.jisc.ac.uk/uploaded_documents/SUMMARY%20FINAL%202005.doc (Accessed 18 /02 06)

Perry, D. (2003)(Handheld Computers (PDAs) in Schools http://www.becta.org.uk/page_documents/research/handhelds.pdf (Accessed 20/06/06)

Roibas, A. C. and Sanchez, I. A. (2002) In *MLearn 2002 Conference*(Eds, Anastopolou, S., Sharples, M. and Vavoula, G.) The University of Birmingham, Birmingham, pp. 53.

Sharples, M. (2000a) *Computers & Education,* **34,** 177-193.

Sharples, M. (2000b)(Disruptive Devices: Personal Technologies and Education http://www.eee.bham.ac.uk/handler/ePapers/disruptive.pdf (Accessed May 10th 2003)

www.expresso.co.uk (2005)(Innovative learning project wins national award for Wolverhampton City Council and Espresso http://www.espresso.co.uk/news/press_releases/051026_wolves_pda.html (Accessed 18/02 2006)

CONSTRUCTING ASSESSMENTS USING TOOLS AND SERVICES (CATS)

Iain Tulloch, James Everett, Rowin Young, Morag Watson and Robin Taylor

Constructing Assessments using Tools and Services (CATS)

Iain Tulloch, James Everett, Rowin Young
Department of Learning Services
The University of Strathclyde
155 George Street
Glasgow
G11RD

Morag Watson and Robin Taylor
Edinburgh University Library
The University of Edinburgh
George Square
Edinburgh
EH8 9LJ

Overview

CATS is a JISC ELF Demonstrator project and represents a collaborative undertaking between the Universities of Strathclyde and Edinburgh. Its goal is to develop a system which returns a complete content-packaged assessment – i.e. a structured set of items - by querying one or more item banks. The main tasks of the project are to:

- analyse and scope the functional requirements of such a system
- create web services to search for, retrieve and aggregate items held in item banks
- utilise and build upon the outputs of two previous JISC ELF projects – SPAID and Discovery Plus (D+)
- consult with assessment domain practitioners

The **SPAID** (Storage and Packaging of Assessment Item Data) system will be used to establish and populate a test item bank. The **D+** system will be enhanced to search for and retrieve assessment items held in one or more item banks. The consultation exercise will canvass practitioners about the types of query they would wish to specify inside a CATS "profile" (see below).

Prototype System Design

Figure 1 represents the high-level web-service architecture of the CATS prototype system. It also illustrates the processing of a request to produce an aggregation of assessment items as a content package. It assumes that at most two item banks will be searched.

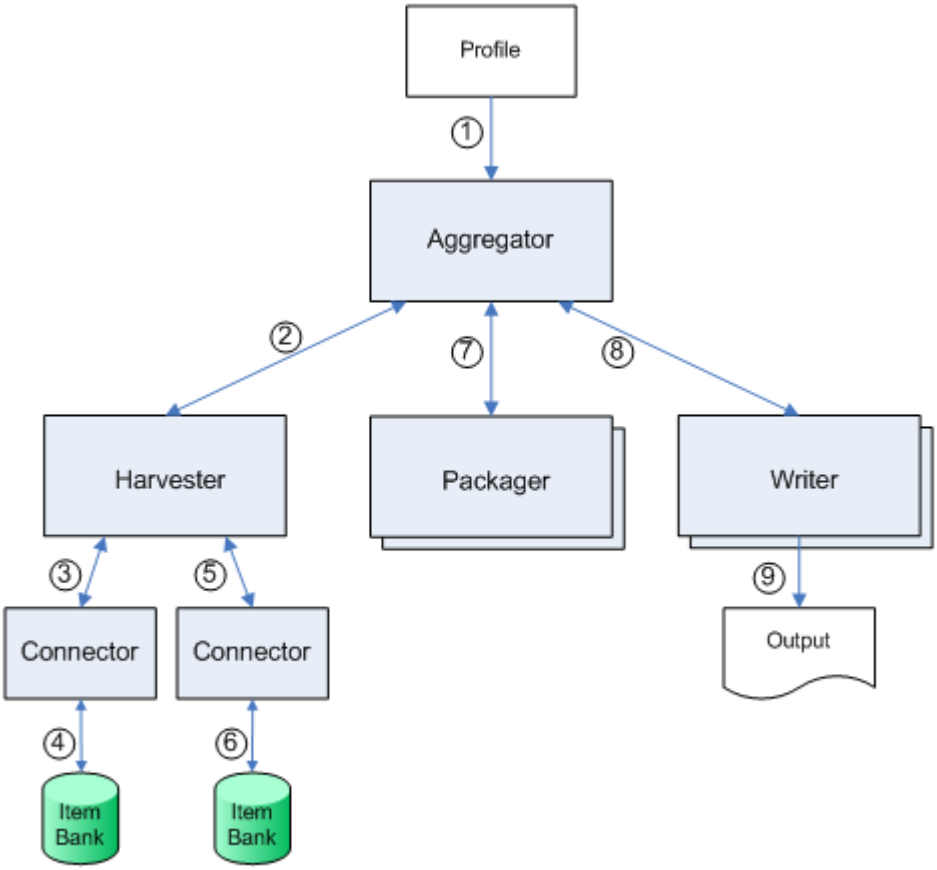


Figure 1: Design of CATS prototype system

The input to the CATS system is a static user-defined "profile" – a parameter file specifying *inter alia*

- the query to execute (e.g. "retrieve 10 geometry items")
- the identifier(s) of the target item bank(s) to search
- the identifier of a content packaging service
- the identifier of a file writing service

The first step is to issue a request to the Aggregator service, passing in a profile (*Step 1*). The Aggregator subsequently calls the Harvester service, supplying the query and the list of target item banks from the profile (*Step 2*).

The Harvester creates an instance of a Connector service for the first target item bank (*Step 3*). A Connector is a D+ based service which searches for and retrieves entire items stored within a specific item bank.

The Connector executes the query on the target item bank; matching items are returned (*Step 4*).

The Harvester may then call another Connector; for example, the number of items returned from the first bank may fall short of the total number requested in the profile. In this case, the Harvester will create an instance of a new

Connector for the next target bank (*Step 5*). This scenario represents invoking, inside the Harvester service, a “collation algorithm” i.e. a prescribed set of steps which resolves a query across the available item banks.

This second Connector instance executes the same query on the second target item bank; again matching items are returned (*Step 6*).

The Aggregator passes all the items returned by the Harvester to the Packager service specified in the profile; the Packager then returns a content package (containing manifest and packaged item(s)) (*Step 7*).

The Aggregator then passes this content package to the Writer service specified in the profile (*Step 8*).

Finally, the Writer outputs the data file(s) for the package (*Step 9*).

Future Enhancements to CATS

We have identified potential enhancements to the current system e.g.

- The prototype has no user-interaction. We believe there are several points during the flow of execution where user intervention would be valuable e.g.
  - allow user to build a search query interactively, using a GUI interface
  - allow user to preview, then retain or reject items returned to the Harvester, and re-query Harvester for more items if desired
- Support a variety of collation algorithms inside the Harvester (e.g. parallel searching of multiple targets)
- Allow specification of output destination e.g. the ability to send a complete assessment directly to an external AMS (Assessment Management System).

Workshop

The workshop will provide a useful opportunity to report work to date and capture feedback on the CATS project.

In this session we will:

- present an overview of the project
- demonstrate the functionality of the CATS prototype
- conduct a group-based discussion covering:
  - the overall practical value of such a system
  - the design approach we have taken
  - use-cases and scenarios
  - possible future enhancements to the system

QUICK WIN OR SLOW BURN? MODELLING UK HE CAA UPTAKE

Bill Warburton

Quick Win or Slow Burn?
Modelling UK HE CAA Uptake

Bill Warburton
University of Southampton
W.I.Warburton@soton.ac.uk

Abstract

The uptake of CAA in UK higher education (HE) on a large scale lags behind the expectations of CAA specialists. A research project was undertaken with the aim of discovering and addressing the underlying reasons for this. The research was conducted according to Strauss and Corbin's (1998) prescription for grounded theory (GT) research. During three years a 200 000 word dataset was compiled from a national survey by questionnaire and interview with tutors, learning technologists, managers and QA staff. This article describes the dual-path theory of CAA uptake that emerged from an analysis of this dataset. Ways in which dual-path theory might be used to understand and improve CAA uptake are proposed.

Quick Wins?

Time pressures on tutors across the sector are well documented (Bull, 1999; Gibbs, Habeshaw and Yorke, 2000) and are often compounded by increasing demand for research output that will raise their profile in the next research assessment exercise (RAE). This promotes a utilitarian approach to assessment activities which prizes quick returns above pedagogic gains or longer term considerations such as an expected reduction in assessment load once a large item bank has been built. CAA was widely acknowledged to offer the potential of productivity gains in terms of more efficient authoring, publication, delivery, marking and reporting, which was summed up by some respondents as an effective reduction in paperwork.

However it also emerged that where unsupported tutors sought these 'quick wins' without investing in preparative activities such as seeking the advice of experienced colleagues or setting up formative exercises and practice quizzes, the degree of risk taken on all at once could be so significant that colleagues were discouraged from using CAA themselves. This effect was prominent in extreme cases such as student data loss during an invigilated examination:

… when the email came round about the [CAA] disaster… some of those colleagues… just went non-linear… how can we possibly have… taken on something which under the most fundamentally obvious things that it had to work under, it fails at the first hurdle? (Tutor AmO5M007)

The effect was less pronounced where the unfavourable outcome was limited to unplanned expenditure of time and effort, for example to recover data or reassure students. Failure to think through the implications of using CAA can have serious implications:

… a CAA had been taken and the results had been distributed to [an inexperienced] tutor, the tutor had given them to someone… who… sent them to an external [examiner], including a detailed breakdown of the item analysis of the assessment, which the tutor didn't understand and hadn't intended to go. So the external [examiner] looked at all this and said 'thank you very much, your test appears to be invalid'. (Learning technologist LtO3M001)

Unintended outcomes of this kind threaten the CAA user's credibility. The increased risk incurred by productivity-driven approaches to CAA applications and the braking effect they have on uptake by colleagues represents an extreme case and is shown in the upper half of the paradigm model (Figure 1). It should be noted that this opening of the assessment process to public scrutiny could be regarded as an unintended consequence of CAA which is seldom included in risk registers. Until recently assessment feedback was rarely given, not least because the examination system was ill equipped to provide it. Therefore participants didn't expect feedback and there was no possibility of a debate about academic standards. Now people know it can be done so they take it for granted, not only for formative and diagnostic use but also for summative assessment as well.

Slow Burn?

Conversely, where tutors aimed primarily for pedagogical improvements they incurred much less risk and the resultant trajectories were characterised by routine use of small scale quizzes with an initial emphasis on low stakes testing such as formative and diagnostic applications. This sometimes progressed towards higher stakes testing on a larger scale.

A staged approach was encouraged by learning technologists who recognised the value for tutors of learning to use complex CAA tools in less critical applications. High stakes applications such as examinations were seen by learning technologists as the final goal of CAA trajectories rather than a starting point. Experienced CAA using tutors agreed.

Staged lower risk trajectories generally produced modest productivity gains and consequently diffusion was steady rather than spectacular. Where tutors emulated this approach, they appeared to do so because they perceived a structured, methodical pattern of practice which would protect their investment in assessment materials and which might yield sustainable if modest productivity gains in the medium to long term.

The reduced risk incurred by pedagogically-driven attitudes to CAA use and the accelerating effect this has on uptake by colleagues is shown in the lower half of the model (Figure 1).

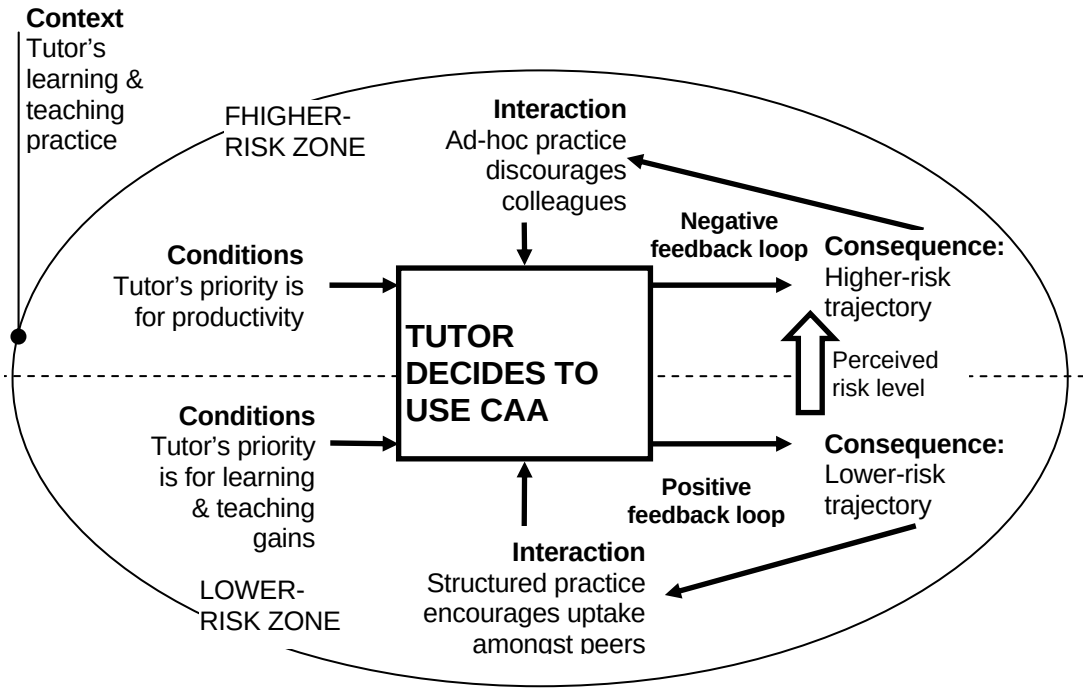


Figure 1. Core dual-path theory of uptake

Internal risk mitigation

In cases where tutors are already experienced, or are supported by experienced colleagues and learning technologists, this constituted a degree of risk mitigation that could shift what would otherwise have been risky CAA practice into a lower risk trajectory. This mitigating action could be taken by CAA users themselves as 'internal' risk mitigation or by learning technologists on their behalf as 'external' risk mitigation.

External risk mitigation

In other cases risk mitigation was performed by learning technologists, who were keenly aware of the underlying fragility of CAA systems ('… the least little thing missed can knock the whole system out' - Learning technologist LtO3F002). An overarching aim of these activities was to make CAA systems easier to use, thus reducing the scope for things to go wrong.

A physical aspect of the risk mitigation that learning technologists undertook was to ensure that the integrity of CAA systems, including associated infrastructures, was beyond reproach. These physical measures were sometimes triggered by problems that occurred during high-stakes use where risky practice had exposed underlying weaknesses such as scalability issues:

… this is its first semester of use and the take-up was so high - so much higher that it led to fairly spectacular problems with it, which… we've now sorted by tuning the system (Learning technologist LtO5M002)

A cultural aspect of risk mitigation by learning technologists was to ensure that appropriate CAA procedures existed and were observed by tutors. CAA policies and procedures were easily overlooked:

… we had an incident this year where one of the lecturers… overlooked a procedure which compromised the exam just beforehand and now they have gone off using the system as a result of that oversight. So even though the procedures were in place and he neglected to do one aspect, it has tarnished his view on [CAA]. (Learning technologist LtN2M003)

Risk mitigating measures of both kinds were taken by learning technologists in a recursive fashion which resulted in a progressively closer fit of mitigation to practice (Harwood and Warburton, 2004).

Strategic Support

The role of strategic support in legitimating CAA was particularly evident in new universities where centralised organisational structures facilitated the promulgation of CAA policies and procedures:

… ultimately we have got one [group of] staff who… filter down all the teaching practices [and] they decide what should [happen] and… it gets validated by them: quality procedures and everything… then things come down from the top and CAA practices are imbedded... (Learning technologist LtN2M003)

This is shown as *institutional validation of existing good practice* and has the direct consequence of increasing uptake by strengthening the remit of the procedural measures put in place by learning technologists. It has the indirect consequence (shown as a dashed line) of increasing uptake by demonstrating the institution's commitment to CAA as a valid tool in the teaching and learning toolkit. The other way in which institutions could drive CAA uptake was by providing a *secure funding* and thereby further validating CAA. This increases uptake by strengthening the physical infrastructure and, by virtue of committing real resources, has the indirect consequence of increasing uptake by demonstrating the institution's commitment to CAA

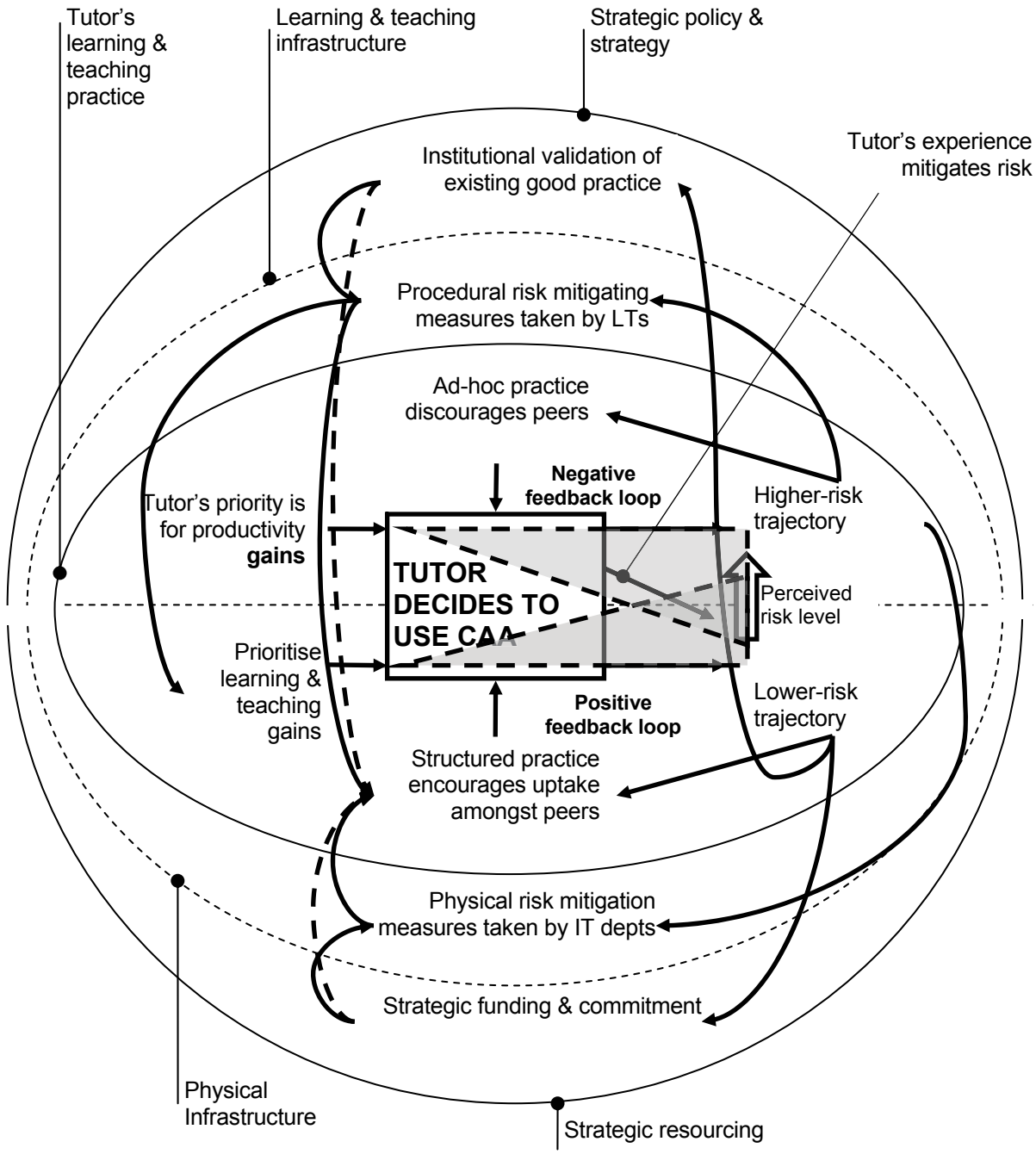


Figure 2- Enhanced dual-path theory showing the influence of strategic support on risk mitigation

The Concentric Shell Model of Uptake

Populating the enhanced dual-path theory with drivers and obstacles identified in specific institutions results in a concentric shell model of uptake which can be used to identify action for optimal uptake (Figure 3).

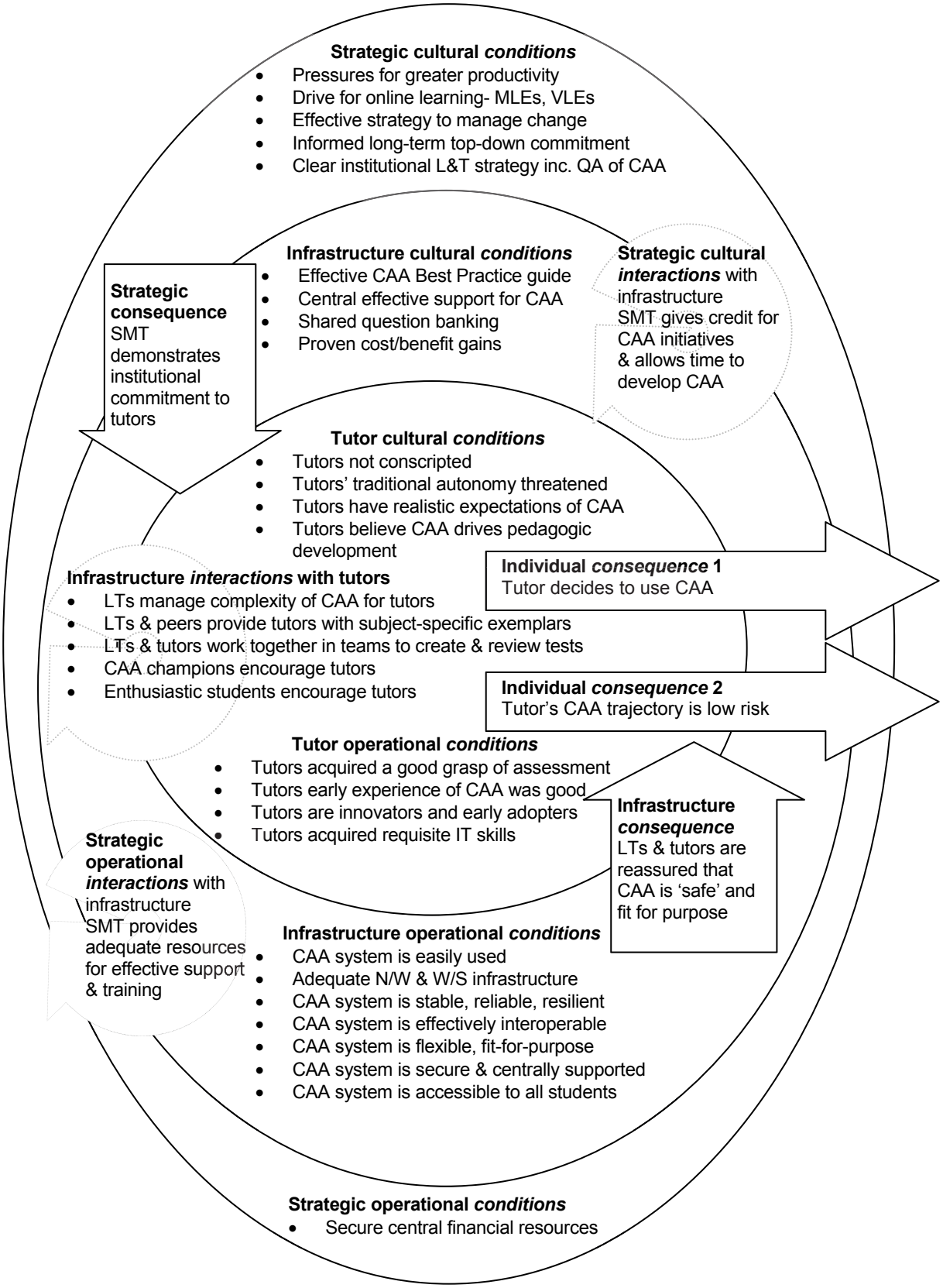


Figure 3- Concentric shell model of CAA uptake showing known drivers

Tutor Trajectories

The pattern of CAA uptake over time at the level of individual tutors - their 'trajectory' - is the fundamental unit which, on the micro scale, underlies institutional uptake on the macro level. A tutor's CAA trajectory differs critically from otherwise similar patterns of technology uptake such as VLE use in that a significant element of risk attends technology-based assessment activities, particularly in credit-bearing assessment.

Individual CAA trajectories can be broadly characterised as high or low risk according to the fashion in which tutors progress towards high stakes assessment. Where uptake proceeds in a planned sequential fashion from testing through formative to low and then high stakes summative testing, small increments of risk are incurred in each step which results typically in a linear low risk trajectory. Where uptake proceeds directly to summative use, large increments of risk may be incurred at once which results typically in a non-linear high risk trajectory. The biggest influences on tutor trajectories were their motives for using CAA. Where the aim was primarily to secure productivity gains the consequence was an ad hoc style of use that resulted in high risk trajectories. Where the aim was primarily to improve learning and teaching practice the consequence was a sustained progression through the different stages of CAA use that resulted in lower risk trajectories (Figure 4).

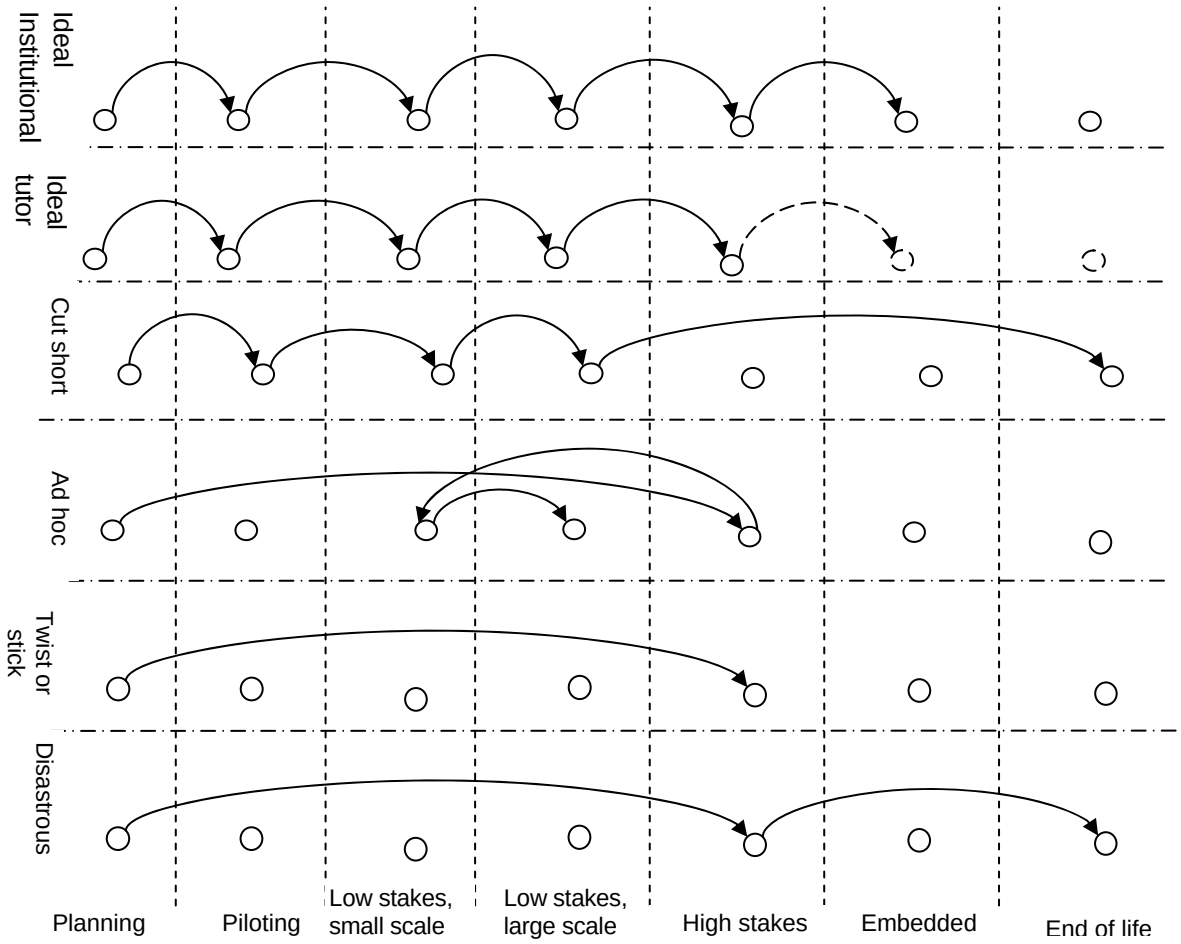


Figure 4- Typical trajectories

Principle Mechanisms Driving CAA Uptake

The principle mechanisms appeared to be sevenfold. They are described in ascending order of scale using the concentric cylinder model of uptake (Figure 5).

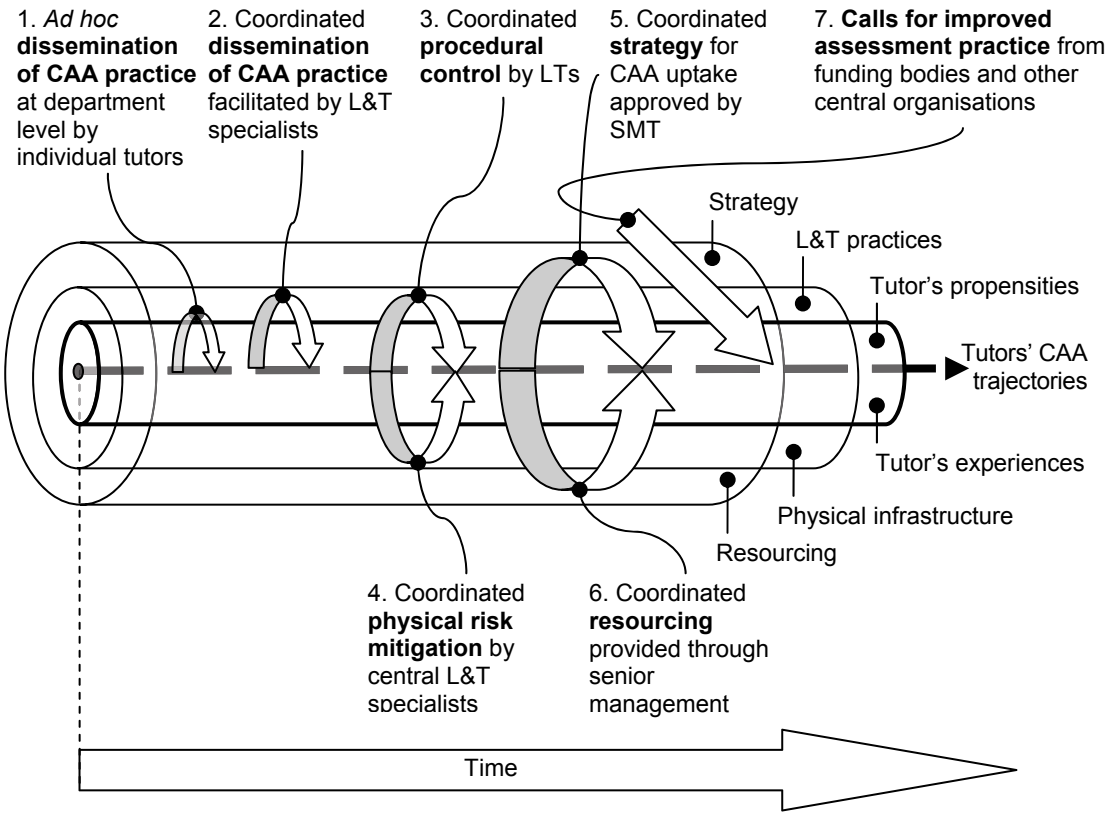


Figure 5 Concentric cylinder model of principle mechanisms driving CAA uptake

It was noted that these mechanisms incur greater latency as they reach higher into the infrastructural and strategic parts of the institution.

1. Ad hoc dissemination of CAA practice at department level

The simplest and most direct form of diffusion is unaided 'word of mouth' dissemination among individual tutors who work together as colleagues. This is recognised by learning technologists and tutors as an effective driver which acts 'horizontally' with respect to other tutors.

2. Coordinated dissemination of CAA practice

One aspect of the model that hinged on mediated support from learning technologists was achieving a 'critical mass' of CAA use. Learning technologists in centralised institutions have a strategic role which permits them to coordinate update by controlling uptake directly from the top down:

3. Coordinated procedural risk mitigation

In some more centralised universities procedural risk mitigation enforces lower risk practice through institutional *fiat*:

4. Coordinated physical risk mitigation by central L&T specialists

Tutors and learning technologists who had experience of high stakes CAA testing were keen to reduce the chance of something going wrong at a critical time by having institutions invest in suitable physical infrastructures.

5. Coordinated strategy for CAA uptake approved by senior management

Having a member of senior management as an advocate for CAA was cited as crucial by experienced learning technologists. Efforts to develop integrated managed learning environments (MLEs) at a strategic level were identified as both an obstacle where absent and a driver where present. The relationship between the uptake of VLEs and of CAA uptake was described as one where neither could advance more than one step beyond the other. Tutors have to make their own logistical arrangements for high-stakes summative tests when institutions do not support CAA examinations via the Examinations Office. This presents an effective obstacle to uptake.

6. Coordinated resourcing provided through senior management

There was clear agreement from learning technologists and tutors about the central importance of centralised support and resourcing:

… when I was at Havenpool, it sort of failed simply because the central services didn't take it on… something about the way it was done without a central team… So there was no central agreement and no institutional drive, so it didn't work, no-one really was sure of who's doing what and why were they doing it anyway, you know? … you need [the institution] to build a solid foundation… (Learning technologist LtN4F001)

7. External influences

Central government funding initiatives may drive uptake by providing an incentive for institutions to implement centralised CAA systems. The pressure from the quality assurance agency (QAA) for more frequent formative feedback should not be underestimated as a driver for uptake at the level of individual tutors:

… there is an awful lot of pressure on teachers … to provide feedback to students... And that's where…[CAA]… is a scalable method of giving feedback to students as they progress through... the QAA are kind of very heavy about [formative feedback] at the moment… the students… go through the semester, they get a semester exam and there's nothing that…could have ever told them how they were doing. (Tutor AmO4M017)

Principle Mechanisms Inhibiting CAA Uptake

The principle mechanisms that emerged from the questionnaire returns and the interviews as inhibiting the uptake of CAA in UK universities were also sevenfold and are described in ascending order of scale. They are depicted below using the concentric cylinder model of uptake (Figure 6).

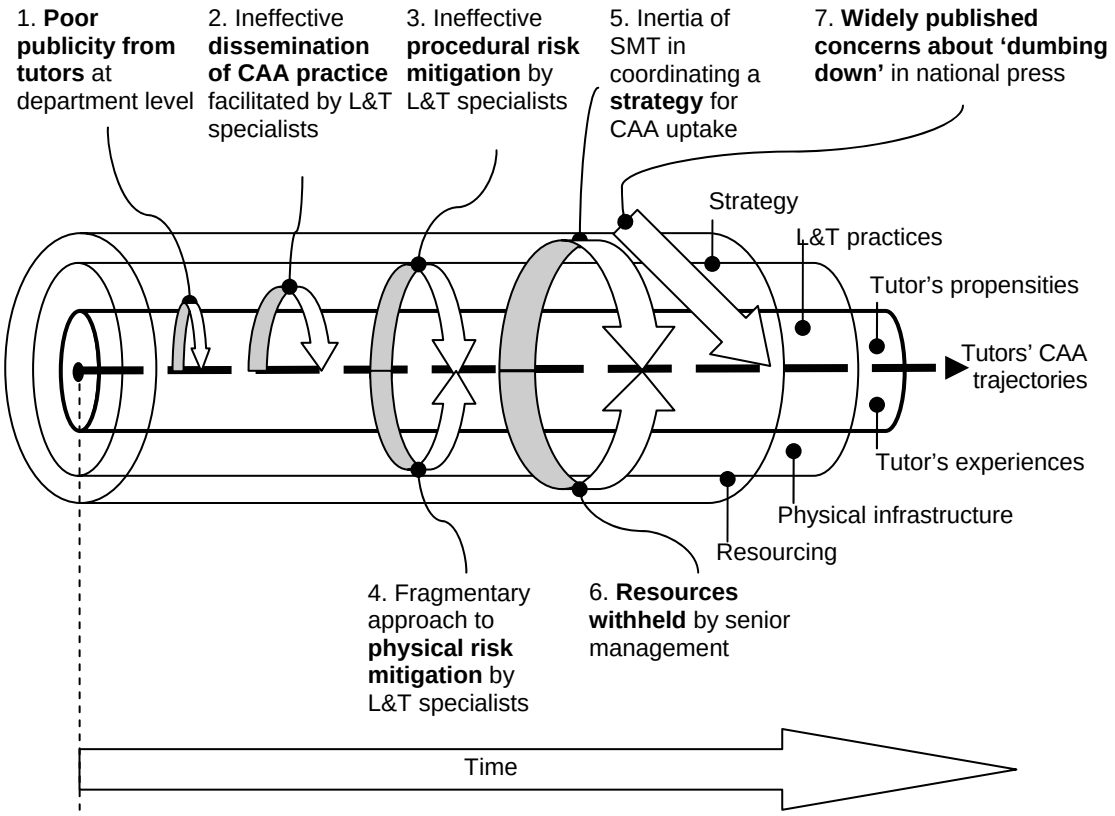


Figure 6- Concentric cylinder model of principle mechanisms inhibiting CAA uptake

1. CAA failures of invigilated tests and fear of these

CAA failures, especially in high stakes invigilated summative tests, have serious consequences for uptake at every level. The consequences are most severe for the tutor because students feel they are under enough pressure without assessment glitches to make things worse. Fears of embarrassment about high-stakes failures resulted in 'confidentiality bubbles' (Harwood, 2002) that restrict diffusion of these events beyond the boundaries of individual academic departments or groups of learning technologists. The basis of this embarrassment appeared to be a perceived threat to the credibility of tutors and departments:

… [we thought] they'd tell us it was our own fault or something… there's that nagging feeling you get that you forgot to do something vital, like did you turn off the gas? (Tutor AsO4F003)

This *under-reporting of CAA failures* contributed to a widespread perception that high-stakes CAA tests were less risky than they really were, which acted

as a driver for uptake particularly among tutors who have naïve understandings of technology:

I think its more of a problem with the staff is their tendency to overestimate their ability to use computers…They think maybe because they can use a wizard in POSH-CAA, that… they're an author for CAA… (Learning technologist LtO3M004)

2. Ineffective dissemination of good CAA practice

CAA uptake is vulnerable to attacks from vociferous critics who may have their own agendas based on perceived threats to a department's credibility:

There's probably a few people [here] who'd love to see one go wrong so they could avoid it, I think and never touch the system again. It's a bit Machiavelli. (Tutor AsO3M002)

The 'quick win' attitude towards CAA is clearly recognisable as a brake on uptake through external examiners' reports to departments:

[external examiners] realise that there are good ways of using it... but there are other staff who see it as a timesaver and therefore do not put as much time into question development and management as could be put in, therefore tests are not as academically testing as could be - so [external examiners] are not as happy... (Tutor AmN3F001)

3. Ineffective procedural risk mitigation

Procedures which do not yet exist, or which are difficult to interpret, constitute an effective obstacle to uptake. Failures to comply with known procedures can have devastating effects on CAA uptake:

…we had an incident this year where one of the lecturers …neglected a procedure which compromised the exam just beforehand and now they have gone off using the system as a result of that oversight. So even though the procedures were in place and he neglected to do one aspect, it has tarnished [their] view on [CAA]... (Learning technologist LtN2M003)

4. Fragmented approach to physical risk mitigation

CAA systems which are not made easy to use are regarded by both tutors and learning technologists as a significant obstacle to uptake:

And I do think you are totally right about the infrastructure and operational conditions and one of the things I've introduced… - well it would take maybe 10 minutes if you were really slick… and in that 10 minutes you could have covered a chapter in the syllabus. So only the really keen ones did it. So I think the infrastructure, yes, is a crucial thing there. Yes, ease of use, that's right, exactly - it is, yes. (Learning technologist AmO5M007)

The difficulty of load-testing CAA systems emerged as a significant obstacle to uptake.

5. Institutional strategy shortfall

The inertia associated with institutions approving CAA applications acts as a brake on innovation by leaving little time for busy tutors to change their practice. As a complement to institutional inertia, one learning technologist cited ongoing organisational change as being itself an obstacle to innovation in assessment:

And it's exactly an inertia of change which is a ridiculous thing to say, but because we're changing we can't do a lot of things. (Learning technologist LtO5M006)

Learning technologists identified failure to implement an overarching strategy at the institutional level as a significant brake on uptake because those wishing to use CAA in summative applications are often obliged to wait for institutions to give permission.

6. Resources withheld by senior management

According to learning technologists, the pace of organisational change was sometimes cited by senior management as a good reason for not investing in institutional CAA infrastructure such as large workstation areas:

What you're talking about is not investing a lot of money in a large, or several large 200-seat computer clusters. I have a sneaking suspicion here that the actual driver behind this is that the University doesn't like spending money. (Learning technologist LtO5M006)

A reluctance on the part of senior management to invest in infrastructure until uptake had increased to the point where it was justified was said to compound the lack of suitable workstation areas as a brake on uptake:

… I've been told that we won't get infrastructure unless we can demonstrate there's a demand. The problem is you can't stimulate the demand unless you can demonstrate there's an infrastructure in which it can work. So its one of these sorts of circular arguments, where it's very difficult to know how it's going to be taken forward. (Learning technologist LtO5M006)

7. Widespread concerns about 'dumbing down'

Fears of 'dumbing down' inhibit uptake by affecting the perceptions of external stakeholders such as employers regarding the use of CAA in HE. This may have discouraged some departments from using CAA:

… external factors… may have a knock on effect for the university if it is using CAA if there a perception by the employers that it's no good and they won't employ people because of this then they might stop using it and switch to more traditional assessment methods. (Learning technologist LtN2M003)

Applications of Dual-path Theory

Three models describing different aspects of uptake emerged from the central dual-path theory. These were the trajectory, concentric shell and concentric cylinder models which could be used both to identify weaknesses in HE institutional practice and to suggest where resources should best be targeted to strengthen uptake. For example, an institutional survey of CAA users and non-users could furnish a register of site-specific obstacles and drivers to populate the concentric shell template. This would illustrate the local balance of existing good CAA practice compared with applications might benefit from mediation. The impact of cumulative institutional hysteresis would be shown by populating a concentric cylinder template with local equivalents of known factors such as an incoherent learning and teaching strategy.

The Contended Notion of 'Successful' Uptake

The uptake of CAA must be considered in the context of 'successful' practice. If a consensus exists that practice across an institution is optimal then there is little incentive to take corrective action. However, stakeholders were found to take different views of this according to their position within the institution. For example, tutors tended to concentrate on completing assessments tasks with maximal efficiency (and minimal student unrest) whilst learning technologists were interested in pedagogic fitness for purpose and extending technical boundaries. The importance of scale emerged as another contentious aspect of uptake (Figure 7).

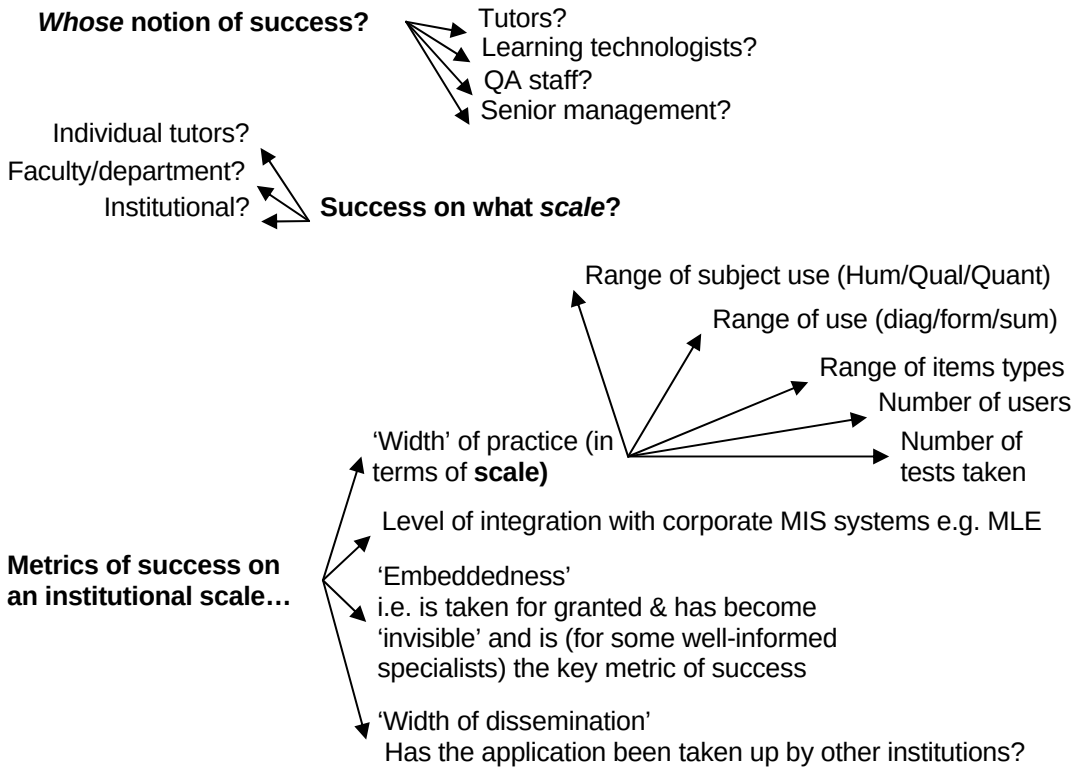


Figure 7- metrics for successful implementation

Discussion

At the level of individual tutors, risk propensity appears to be a good predictor of CAA trajectory type and could be used to direct support resources where they might be used most effectively to mitigate risky practice. Trajectories seem to be good descriptors of CAA uptake patterns and provide an effective and concise way of characterising existing and future practice. Metrics for 'good' CAA practice are admittedly contentious but efforts must be made to establish reference points which are recognisably grounded in wider communities of practice. The crude distinction drawn here between the 'quick win' and 'slow burn' patterns of uptake could be taken as the simplest possible way of differentiating different patterns of CAA practice. It might be argued that a lack of clear descriptors has until now contributed to the difficulty of agreeing common reference points for characterising uptake.

This paper described the development of a grounded theory of CAA uptake in UK HE institutions, known as the 'dual-path' theory. Three models developed from this theory can be used to understand CAA practice at the levels of individual tutors, infrastructure and entire institutions. These models can be used to identify weaknesses in HE institutional practice and to suggest where resources might be committed to optimise uptake. Notions of 'successful' uptake are contentious due to differences in stakeholder perspective.

References

HARWOOD, I. (2002) *Developing Scenarios for Post-Merger and Acquisition Integration: A Grounded Theory of Risk Bartering*. Unpublished PhD thesis, University of Southampton.

HARWOOD, I. & WARBURTON, W. (2004) Thinking the Unthinkable: Using Project Risk Management when Introducing Computer-assisted Assessments. IN: DANSON, M., ed., *Proceedings of 8th International CAA Conference*. Loughborough, University of Loughborough.

WARBURTON, W. & CONOLE, G. (2005) Wither e-Assessment? IN: DANSON, M., ed., *Proceedings of 9th International CAA Conference*. Loughborough, University of Loughborough.

AN INVESTIGATION OF THE RESPONSE TIME FOR MATHS ITEMS IN A COMPUTER ADAPTIVE TEST

Chris Wheadon and Qingping He

An Investigation of the Response Time for Maths Items in a Computer Adaptive Test

Chris Wheadon and Qingping He
CEM Centre
Durham University
UK

Chris.Wheadon@cem.dur.ac.uk
Qingping.He@cem.dur.ac.uk

Abstract

An important advantage of computer based testing over conventional paper and pencil based testing is that the response time to items from test takers can be accurately recorded for subsequent analysis. This study investigates the response time for maths items in a computer adaptive test designed as a baseline assessment for pupils aged from 11 to 18 in the UK. The results showed that the response time for all the items in the test generally increases with item difficulty, although significant variability exists. The item difficulty levels and the age and ability of test takers have significant influence on item response time.

Keywords

Item Response Theory, Computer Adaptive Testing, Item Response Time.

Introduction

Information and computing technology (ICT) has been widely used in education at various levels to assist learning in education organisations, and computer based testing (CBT) is becoming increasingly important as an assessment tool (e.g Tymms, 2001; Gardner *et al*, 2002; Ashton *et al*., 2003; Russell *et al*., 2003; Tymms *et al*., 2004; He and Tymms, 2005). CBT can gather more information than conventional paper-and-pencil testing. For example, it is possible to record the time a person takes to answer a specific item in a computer-based test. Of the computerised testing procedures currently in use, computer adaptive testing (CAT) has attracted particular attention in recent years (see Lilley and Barker, 2003; He and Tymms, 2005). Most computer adaptive testing systems are based on the implementation of an Item Response Theory (IRT) model, which generally assumes that, given a test and examinee sample, the overall performance of an examinee is determined by his/her ability and the characteristics of the test items (see, for example, Hambleton and Swaminathan, 1983; Masters and Keeves, 1999; Tymms, 2001; Wang and Kolen, 2001; Tonidandel *et al.*, 2002; Lilley and Barker 2003; He and Tymms, 2005). In a computer adaptive test, for a

particular examinee, the items, drawn from an item bank containing items that have been calibrated using an IRT model (i.e. item statistics such as item difficulty and discrimination power have been estimated using an IRT model), are targeted at his/her ability level, and each individual will therefore answer a different set of items.

The study of item response time is important for understanding the physiological behaviour of test takers during the testing process, which is essential for creating effective items and tests that can provide more accurate educational measurements. A number of researchers have conducted work in this area (e.g. Hornke, 2000; Chiu and Bejar, 2001; Bridgeman and Cline, 2004; Moshinsky and Rapp, 2004; Chang *et al*. 2005). In the study undertaken by Chang *et al*. (2005), the authors found that higher ability students showed persistence with test items irrespective of item difficulty and generally spent more time on items than lower ability students, while work by Moshinsky and Rapp (2004) on a high-stake test used for undergraduate admissions in Israel indicated that: more difficult items generally take more time to answer than easier items and that more able students take less time to answer items incorrectly than less able students.

This paper reports results from an investigation of the response time to the items in an adaptive test based on data collected from over 100,000 students, and attention has been focused on studying the effects of item difficulty, and the age and ability of test takers.

The Computer Adaptive Baseline Test

The Curriculum, Evaluation and Management (CEM) Centre at Durham University has been conducting baseline assessments on primary, secondary and post-sixteen students through the administration of paper-and-pencil based tests and questionnaires via a number of performance indicator related research projects, including the Performance Indicators in Primary Schools (PIPS) project, the Middle Years Information System (MidYIS, for Year 7 students aged from 11-12) project, the Year 11 Information System (Yellis, for Year 10 students aged from 15-16) project and the A Level Information System (Alis for Year 12 students aged from 17-18) project (see Fitz-Gibbon, 1997). The baseline data are then linked to students' subsequent academic performance in order to provide value added information for schools to undertake self-evaluation and management. In view of the relatively good IT facilities available today in schools, a two-part computer adaptive test has been developed as an alternative to the conventional paper-and-pencil baseline tests for the three secondary projects (MidYIS, Yellis and Alis). The adaptive test includes an adaptive maths test and an adaptive English vocabulary test. This computer adaptive baseline testing (CABT) system comprises a calibrated English vocabulary item bank, a calibrated maths item bank, and an item display and recording system for displaying items to students and recording responses. The calibrated item banks, in which the item difficulty varies across a wide range, were established by administrating a series of tests to students of various ages and the embedding of common items in the tests, analysis of test results using the Rasch model (see Rasch, 1960; Wright and Stone, 1979), and the equating of the tests using common

items. In total there are over 500 vocabulary items in the vocabulary item bank and over 500 maths items in the maths item bank. Effort has been made to make the items content-independent to each other when creating the maths items. Testing is delivered through the Web or from the school's local network. As the items in the item banks cover a wide difficulty range, all three projects use the same adaptive tests with different starting item difficulty to gather baseline information. This has avoided the need to develop separate tests for individual projects. The present study will focus on the items contained in the adaptive maths test.

Results and Discussions

The Effect of Item Difficulty

Theoretical models and empirical evidence suggest that there is a positive correlation between item difficulty and response time (e.g. Moshinsky and Rapp, 2004). This relationship is corroborated to some extent in the present study as shown by Figure 1, which plots the response time against item difficulty for all the items in the adaptive maths test taken by year 12 students. However, significant variability in the mean response times at all levels of item difficulty exists. The information presented in Figure 1 will be useful for constructing more efficient tests by using less time-consuming items across a range of difficulty levels.

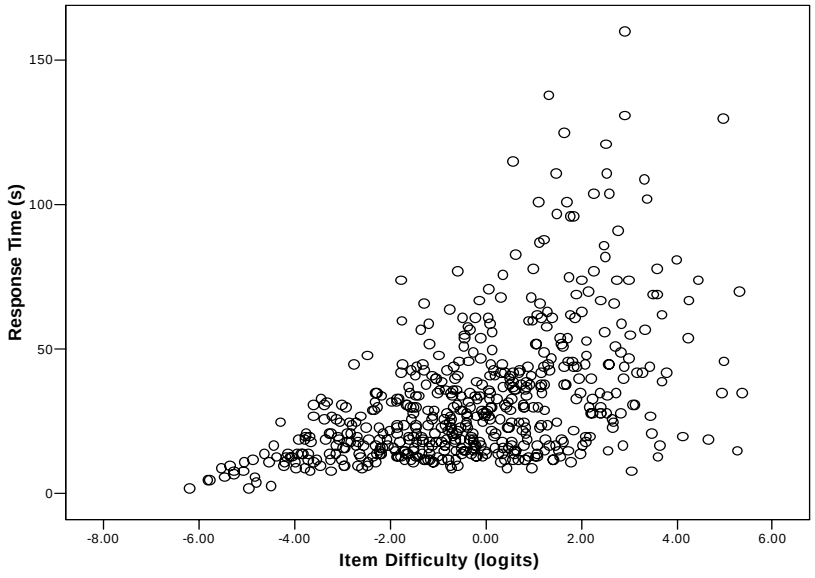


Figure 1 The distribution of response time against item difficulty for Year 12 students

The Effect of Age Groups

As the CABT is undertaken by a large number of students in years 7, 10 and 12, comparisons can be made across age groups. It should be noted, however that the sample size for year 7 decreases as the items become more difficult and the sample size for year 12 increases as the items become more

difficult. As an example, a selection of items from the central difficulty range of the maths item bank have been used for comparison, and Figure 2 shows the distribution of response time for different year groups. Figure 2 shows that Year 7 students seem to spend longer than the other year groups on the easier items than on the more difficult items. This may, however, be due to reduced sample size on the more difficult items. It is clear from Figure 2 that the there is generally a positive correlation in the response time between the groups for the selected items, although the response time varies substantially between the items.

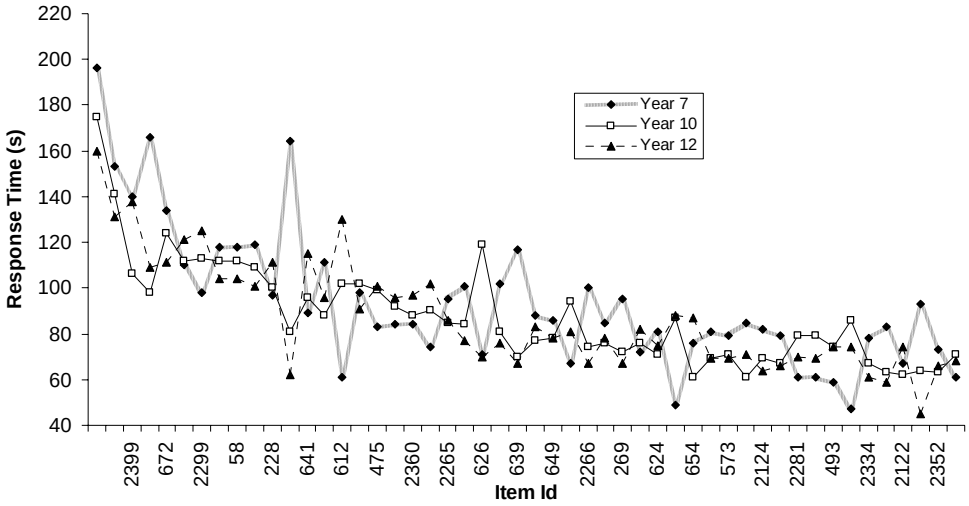


Figure 2 The distribution of response time against year group for a selection of items

Performance Time and Response Accuracy

Figure 3 shows the mean response time by response accuracy across a selection of items. It is clear from Figure 3 that for specific items the mean response time for a correct answer can be greater or less than that for an incorrect answer but there is generally a positive correlation between the two. The overall average response time for correct answers for these items is greater than that for incorrect answers. This is in contrast to the findings from Moshinsky and Rapp (2004). In their study, the authors find that the time reflected in correct responses is less than the time invested in incorrect responses. This contradiction may result from the difference in the nature of the tests being studied: the CABT test is a low-stakes curriculum-free test; the Psychometric Entrance Test is a high-stakes university admissions test. The content domains tested and the age of the test takers may also have contributed to this contradiction.

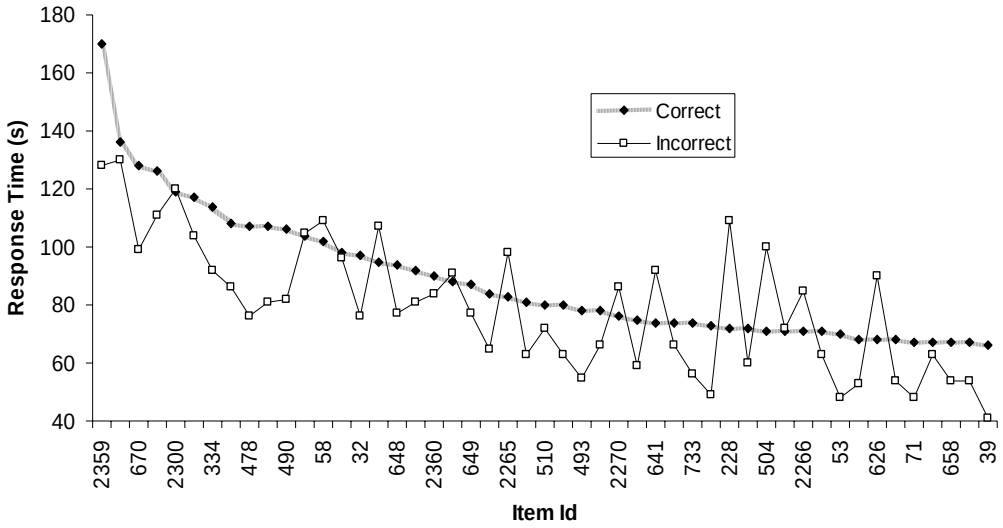


Figure 3 The distribution of the mean response time against response accuracy

The Effect of Ability and Age of Test Takers

Moshinsky and Rapp's (2004) examination of response time found that more able examinees tend to be faster than less able examinees, which is especially true when able examinees know the correct answer. This relationship is diminished when examinees do not know the correct answer, thus the time difference between correct and incorrect answers tends to increase with ability. This was seen to be consistent with the finding that mental ability and mental speed are correlated (see Thissen, 1983). As an adaptive test presents items to candidates that are commensurate with their ability the relationship between time taken on the test and ability is confounded by the difficulty of the items presented to candidates: more able candidates are presented with more difficult items that take longer to solve. Due to the random element of item selection in an adaptive test, and the time it takes for a test to converge on a final estimate of ability, every item is taken by a reasonably wide ability range. The mean ability of 1537 students who took Item 2359 with a difficulty of 2.9 logits, for example, was 2.3 logits, with a minimum of –2.7 logits, a maximum of 8.9 logits and a standard deviation of 1.7 logits. It is therefore possible to analyse the response time of students on individual items which removes any confound with item difficulty. As the CABT is taken by students from the age of 11 to the age of 18 it is furthermore possible to examine the interaction effect of age on performance. As in the study by Moshinsky and Rapp (2004) the correct and incorrect answers are examined separately due to the influence accuracy has on response time.

Three items were chosen for detailed investigation of the effect of age and ability of test takers on response time. These questions, which require a fair amount of time to answer, were chosen from different levels of the difficulty range. The contents of the three items are listed below.

Q.631 Understanding a simple algebraic relationship. Difficulty: -0.6 logits.

The table represents a relationship between x and y. What is the missing number in the table?

X	Y
2	5
3	7
4	?
7	15

a) 9 b) 10 c) 11 d) 12 e) 13

Q.490 Comparing two fractions. Difficulty: 1.1 logits.

Compare the two expressions:

Expression A: $\frac{14+15+16+17+18}{5}$

Expression B: $\frac{17+18+19+20}{4}$

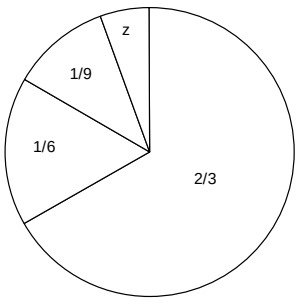
Expression A is greater than expression B

Expression B is greater than expression A

The expressions are equal

Q.2359 Reading a pie chart, working with fractions. Difficulty: 2.9 logits.

The pie chart represents the different colours of cars in Albert Street. If there are 144 cars in total, how many are blue (segment z)?



Free response answer.

Figures 4 to 6 show the relationship between ability and scaled response time (defined as the natural logarithm of the actual response time in seconds) for each year group for the selected items. Care must be taken in interpreting

these graphs, however, for a number of reasons. Response time, as noted by Moshinsky and Rapp (2004) tends to be positively skewed. Natural logarithm and cube root transformations make the distribution more symmetrical, but generally the data is unsuitable for parametric tests. Splitting the Year Groups by ability band furthermore results in uneven sample sizes and heterogeneous variance.

Item 631: An easy item

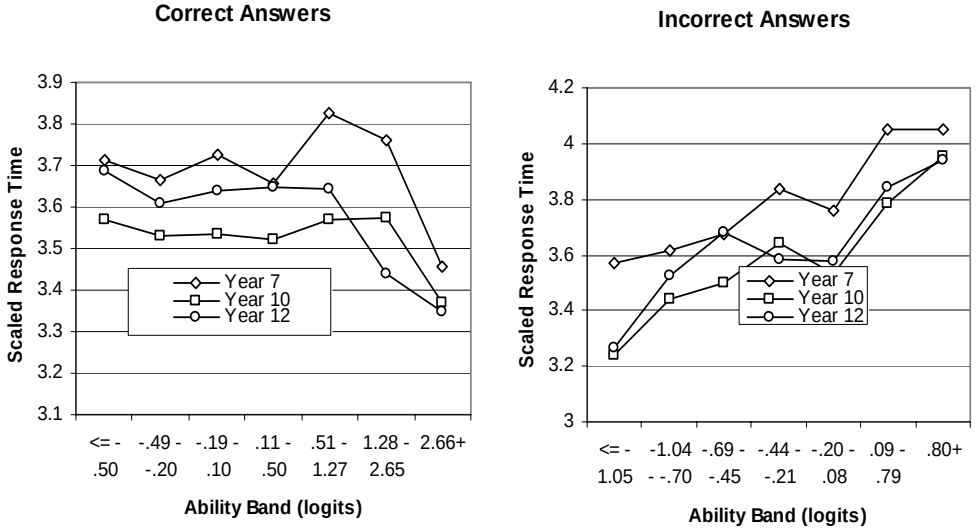


Figure 4 Distribution of scaled response time by ability for correct and incorrect answers on item 631

Figure 4 shows the relationship between response time and ability of test takers (banded) for Q631, which is a relatively easy item. Figure 4 replicates Moshinsky and Rapp's (2004) finding that response time is negatively correlated with ability when the item is answered correctly ($r = -.10$ $p<.001$). Moshinsky and Rapp's (2004) finding that incorrect answers are not correlated with ability, however, are contradicted by the positive correlation ($r = 0.17$ $p<.001$) between ability and response time when the item is answered incorrectly from our study.

For correct answers, response time was significantly affected by Year Group ($H(2)=69.1$, $p<.001$) with Jonckheere's test revealing a significant linear trend in the data, $J = 1852769$, $z = -7.26$, $r = -.012$. Students in year 7 generally took longer to answer this item than students from other year groups. Post hoc Mann-Whitney tests of the difference between the three year groups for correct answers revealed a significant difference between years 7 and 10 ($U=608416$, $r= -.15$) and between years 7 and 12 ($U=457640$, $r=-.16$), but not between 10 and 12. (the critical value for significance was set at .0167 after application of the Bonferroni correction).

Item 490: A item of medium difficulty

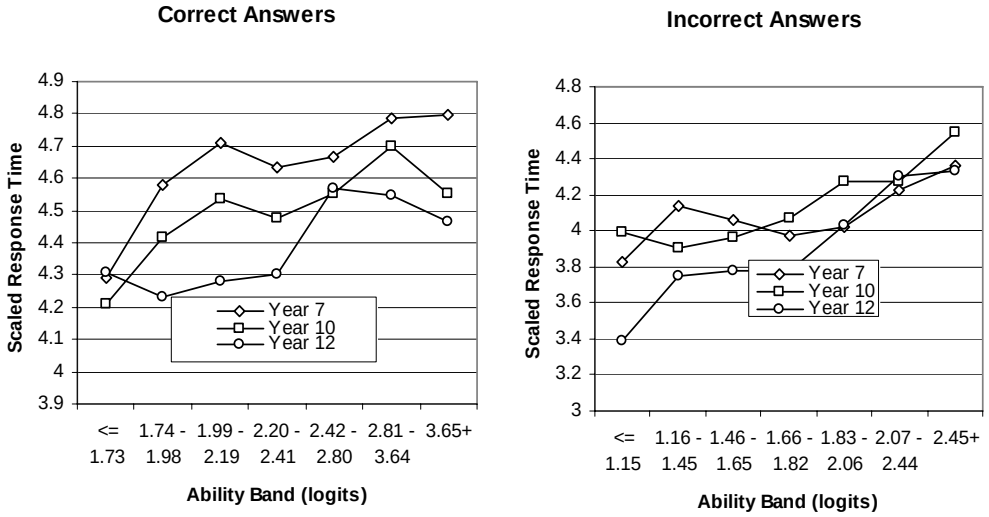


Figure 5 Distribution of scaled response time by ability for correct and incorrect answers on item 490

Figure 5 shows the relationship between response time and ability of test takers (banded) for Q490, which is a medium difficulty item. Figure 5 shows no correlation between time taken and ability for correct answers; contradicting Moshinsky and Rapp's finding that response time is negatively correlated with ability when the item is answered correctly. Moshinsky and Rapp's finding that incorrect answers are not correlated with ability are also contradicted by the positive correlation ($r = 0.14$ $p<.001$) between ability and response time when the item is answered incorrectly.

Once again, for correct answers, response time was significantly affected by Year Group ($H(2)=18.2$, $p<.001$) and Jonckheere's test revealed a significant trend in the data: as year group increases, the median response time decreases, $J=736954$, $z=-4.3$, $r=.09$. Post hoc Mann-Whitney tests of the difference for correct answers between the three year groups revealed no significant difference between years 7 and 10, but a significant difference between Years 7 and 12 ($U=93236$, $r= -.09$) and between years 10 and 12 ($U=547821$, $r= -.07$) with the critical value for significance set at .0167 after application of the Bonferroni correction.

Item 2359: An item of high difficulty

Figure 6 shows the relationship between response time and ability of test takers (banded) for Q2359, which is the most difficult item of the three. Figure 6 replicates Moshinsky and Rapp's (2004) finding that response time is negatively correlated with ability when the item is answered correctly (r -.26 $p<.001$). Moshinsky and Rapp's finding that incorrect answers are not correlated with ability is contradicted by the positive correlation ($r = 0.1$

p<.001) between ability and response time when the item is answered incorrectly.

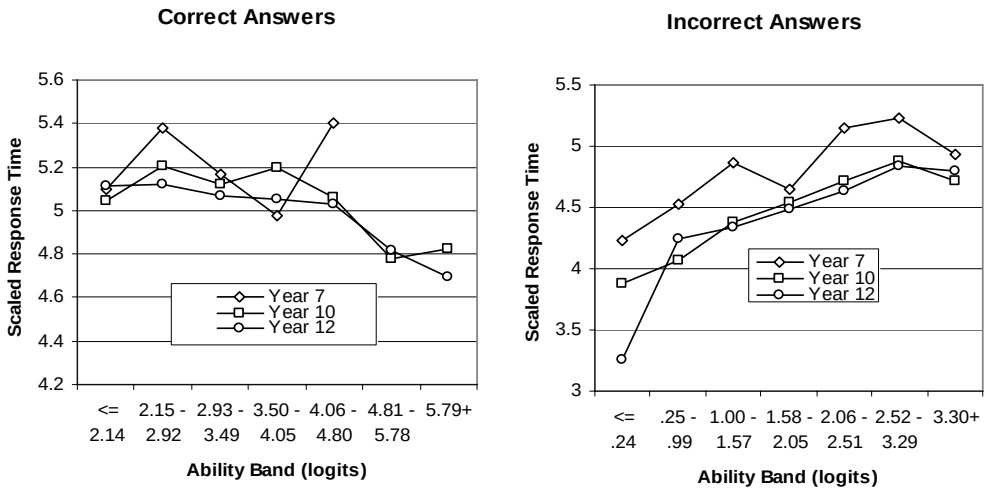


Figure 6 Distribution of scaled response time by ability for correct and incorrect answers on item 2359

Once again, for correct answers, response time was significantly affected by Year Group ($H(2)=11.7$, $p=.003$) and Jonckheere's test revealed a significant trend in the data: as year group increases, the median response time decreases, $J=20392$, $z=-3.4$, $r=-.17$. Post hoc Mann-Whitney tests of the difference for correct answers between the three year groups revealed a significant difference only between years 7 and 12 ($U=4751$, $r=0.2$) with the critical value for significance set at .0167 after application of the Bonferroni correction.

Perspectives

The use of computer based testing, including computer adaptive testing, represents an important advance in educational assessments. An important advantage of CBT is that it can record the time spent by test takers on specific test items which can be used for studying their behaviour during the testing process. Information gathered on item response time is very important for creating effective items and constructing effective tests to provide more accurate educational measurements. Results from this study indicate that the item response time is influenced by a range of factors, including the content domain and difficulty level of the items, and the age and ability of the test takers. Significant variation of response time exists between items and between students with different age and ability.

As the CABT employed in the current study represents a low-stake non-curriculum baseline test, the results obtained can be viewed as an complement to Moshinsky and Rapp's (2004) findings on a high-stakes adaptive test. Our results contradict their finding that there is no correlation between the time taken on items answered incorrectly with ability. The items in the CABT are not curriculum based and often presented in novel ways to the students. Thus it seems that able students persevere for longer trying to

manipulate the item into a form they recognise. The positive correlation between time taken and ability for correct answers found by Moshinsky and Rapp (2004) is not always corroborated. It was most pronounced on the most difficult item where a medium effect size suggested it was an important factor in the response time. This item is most similar to the power items considered in Moshinsky and Rapp's (2004) study, requiring several logical steps, whereas the other items can be answered more quickly in less logical steps. This study furthermore considers the relationship between age and response time. The size of the effect seems to depend on the particular item involved.

While there is a positive correlation between response time and item difficulty, the variation is large. From a technical perspective this offers the opportunity to make the test more efficient, as difficult items that can be answered quickly can be retained in the item bank at the expense of difficult items that take longer to answer. It is not the case that all difficult items are time consuming.

Contrary to Moshinky and Rapp's (2004) findings there does not seem to be a stable relationship between performance time and response accuracy. This may be due to the different levels of familiarity that candidates have on the items.

Further study will involve the use of the response time obtained for individual items contained in the maths item bank of the CABT to design effective tests by selecting items requiring less time to answer at different difficulty levels to investigate the effect of such tests on the accuracy of student ability measurement.

Acknowledgements

The authors would like to thank Michael Cuthbertson, Daniel Bennett, Frank Bell and Peter Clark for their help with data collection.

References

Ashton, H.S., D.K. Scholfiled and S.C. Woodger (2003) Piloting summative Web assessment in secondary education. *2003 CAA Conference Proceedings*: 19-29. Loughborough University, UK.

Bridgeman, B. and F. Cline (2004) Effects of differentially time consuming tests on computer-adaptive test scores. J*ournal of Educational Measurement* 41: 137-148.

Chang, S., B. Plake and A. Ferdous (2005) Response times for correct and incorrect item responses on computerized adaptive tests. Paper presented at the Annual Meeting of the American Educational Research Association, Montréal, Canada.

Chiu, C. and I. Bejiar (2001) An empirical evaluation on the quantitative section of the computer-development and delivered GRE: generalizability analysis of response time and test score. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, USA.

Fitz-Gibbon, C. T. (1997) The Value Added National Project: Final Report *Feasibility studies for a national system of Value Added indicator*. School Curriculum and Assessment Authority, UK

Gardner, L., D. Sheridan and D. White (2002) A Web-based learning and assessment system to support flexible education. *Journal of Computer Assisted Learning* 18: 125-136

Hambleton R. and H. Swaminathan (1983) *Item response theory: Principles and applications*. The Netherlands: Kluwer-Nijoff.

He, Q. & Tymms, P.B. (2005). A computer-assisted test design and diagnosis system for use by classroom teachers*. Journal of Computer Asssted Learning* 21: 419-429.

Hornke, L. (2000) Item response times in computerized adaptive testing. *Psicologica* 21: 175-189.

Lilley, M. and T. Barker (2003) An evaluation of a Computer Adaptive Test in a UK university context. *2003 CAA Conference Proceedings*: 171-182. Loughborough University, UK.

Masters G. and J. Keeves (1999) *Advances in measurement in educational research and assessment*. The Netherlands: Elsevier Science.

Moshinsky, A. and J. Rapp (2004). Performance Time on an Adaptive Power Test. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, USA.

Rasch G. (1960) *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Denmark Paedagogiske Institute.

Russell, M., A. Goldberg and K. O'Connor (2003) Computer-based testing and validity: a look back into the future. *Assessment in Education: Principles, Policy and Practice* 10: 279-293.

Tonidandel, S., M.A. Quiñones and A.A. Adams (2002) Computer-adaptive testing: the impact of test characteristics on perceived performance and test takers' reactions. *The Journal of Applied Psychology* 87: 320-332.

Tymms, P.B. (2001) The development of a computer-adaptive assessment in the early years. *Educational and Child Psychology* 18: 20-30.

Tymms, P.B., C. Merrell, and P. Jones (2004) Using baseline assessment data to make international comparisons. *British Educational Research Journal* (in press).

Wang T. and M.J. Kolen (2001) Evaluating Comparability in Computerized Adaptive Testing: Issues, Criteria and an Example. *Journal of Educational Measurement* 38: 19-49.

Wright, B. D. and M.H. Stone (1979) *Best test design.* Chicago, IL: MESA Press.

INTELLIGENT PENS, PAPER AND INK

Gillian Whitehouse

Intelligent Pens, Paper and Ink

Gillian Whitehouse
Edexcel
190 High Holborn
London
WC1V 7BH
020 7190 4314
gillian.whitehouse@edexcel.org.uk

Abstract

PaperWorks is an EU project concerned with providing distinctive ways of interleaving paper documents with digital materials. The project focuses on developing a core technology for interlinking established content in paper and electronic domains. This is made possible through a non-obtrusive pattern on the paper that allows users to interrelate content with associated digital information. PaperWorks also involves innovative developments in the production of novel substrates, inks, reading devices and the integration of software and communication resources as well as requiring an adaptable information architecture. This is all supported by innovative research to develop support for authoring information and associated links.
Edexcel / Pearson Education have now been working with this project group for 18 months and are developing a method of linking these technologies to both summative and formative assessment processes. The project also involves an analysis of the development and capture of creative and problem solving processes.
Edexcel / Pearson also have a keen interest in developing the links between assessment and learning and through this project are able to demonstrate a variety of potential opportunities that the technology can provide to foster and nurture these links.

Project Partners and Acknowledgements

King's College London, UK (co-ordinator)
Acreo AB, Sweden
Anoto AB, Sweden
ArjoWiggins SAS, France
Brunel University, UK
ETH Zurich, Switzerland
Malmo University, Sweden
Pearson Education, UK
The Technology Education Research Unit (TERU) at Goldsmiths College-London

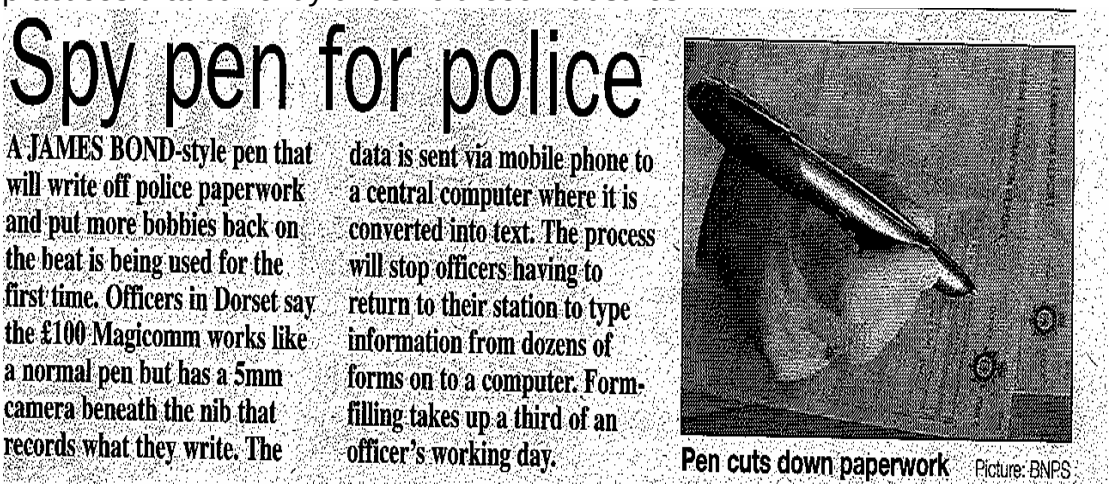
Background

Despite the wide-ranging recognition that paper remains a pervasive resource for human communication and collaboration, there has been uncertain progress in developing technologies to bridge the paper digital divide. This paper revisits the long-standing interest in Computer Supported Co-operative Work (CSCW) with paper, and looks at ways which will enable people to create affinities between material documents and digital resources. An example of this could be enhancing an educational book associated with a television series. Such a book could be augmented to enable the reader to point to pictures or text on the page and gain associated information - video clips and the like - on a workstation, a PDA or television set.
Studies in this area have discovered over and over again one remarkable fact - despite the pervasiveness of new technologies, accompanied in many cases by management's attempt to reshape traditional practice and procedure i.e. the paperless office, paper remained and remains a critical feature of work and collaboration. Many of the examples are well known; - the paper timetable in London Underground ; the traditional medical record in primary health care; (fig. 1) the tickets in financial dealing rooms ; the documents reviewed by lawyers; and so on.

Figure 1.

Paper allows for collaboration because it is mobile, portable between different spaces and regions; it can not only be relocated and juxtaposed with other objects and artefacts, but is micro-mobile, it can be positioned in delicate ways to support mutual access and collaboration. Paper is annotated in *ad hoc* and contingent ways; people can recognise those annotations, track their development and often recognise who has done what. Paper retains a persistent form and preserves the layout and character of art work that is produced on its surface; it can be pictured, memorised, and navigated, even scanned, with ease.
These characteristics and many more not only support complex individual activities but provide a firm foundation to many forms of collaboration, be it synchronous or asynchronous, co-located or distributed. Paper has provided a critical resource to enable people to use technologies, including conventional information systems. Paper is used alongside digital technologies and people spend much time and effort creating, sustaining and transforming the relationship between paper documents and digital resources. Students, teachers, journalists and the like edit text on paper and transpose those corrections to digital copy, architects modify paper plans and integrate those changes in the CAD system, administrators litter their workstations with reminders, diary notes and the like, and booking clerks laboriously write down

the details of your travel arrangements before trying to enter the information into a system. Paper is not just an independent resource that somehow has continued to survive despite attempts to remove it, but rather is an integral feature of using new technologies. Bearing this in mind t is somewhat surprising that relatively little effort has been devoted to enhance the relationship between paper and the digital. If, as seems possible, we can begin to provide people with the ability to access and create links between these resources it will have profound implications for the production and presentation of both. Publishers and some types of broadcasters who produce textual materials alongside programmes would need to rethink and reconfigure the ways in which they structure and present content, and in turn this would have an important impact on the organisational arrangements and practices that currently underlie these industries.



Spy pen for police

A JAMES BOND-style pen that will write off police paperwork and put more bobbies back on the beat is being used for the first time. Officers in Dorset say the £100 Magicomm works like a normal pen but has a 5mm camera beneath the nib that records what they write. The data is sent via mobile phone to a central computer where it is converted into text. The process will stop officers having to return to their station to type information from dozens of forms on to a computer. Form-filling takes up a third of an officer's working day.

Pen cuts down paperwork Picture: BNPS

Aims and Objectives of the PaperWorks Project

The project aims to provide people with new forms of functionality in everyday environments through seemingly mundane artefacts.
The project aims to:

- develop robust, reliable and usable solutions that enable people to access, create and use links between paper and digital resources;
- identify and support applications that enable professional and lay content providers to exploit augmented paper solutions;
- develop innovative hardware and software technologies and techniques that integrate different approaches to augmenting paper.

Description of Work

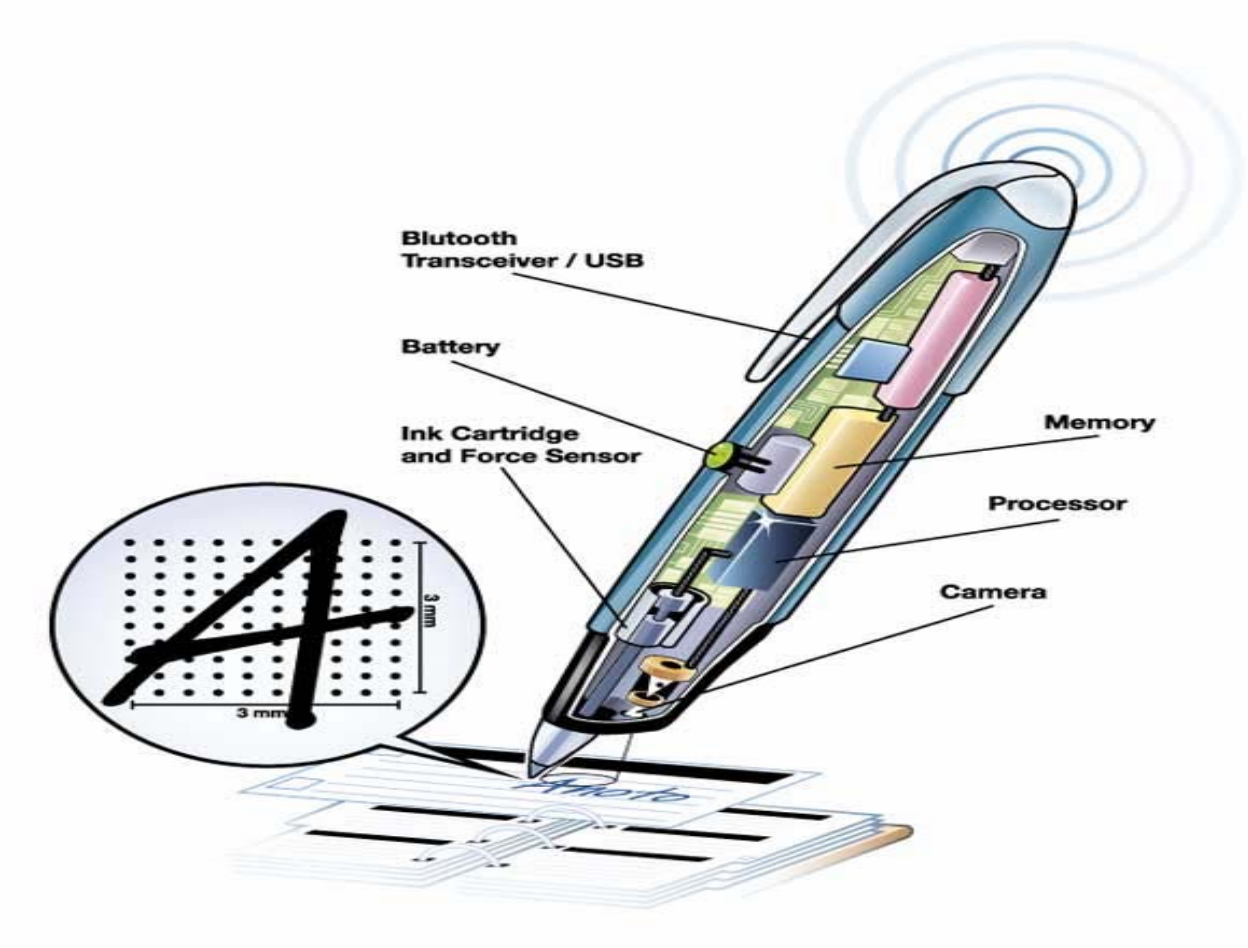
The project undertakes a range of technical, design and empirical activities in order to develop a robust augmented paper solution that could be integrated with a published product. The project is developing:

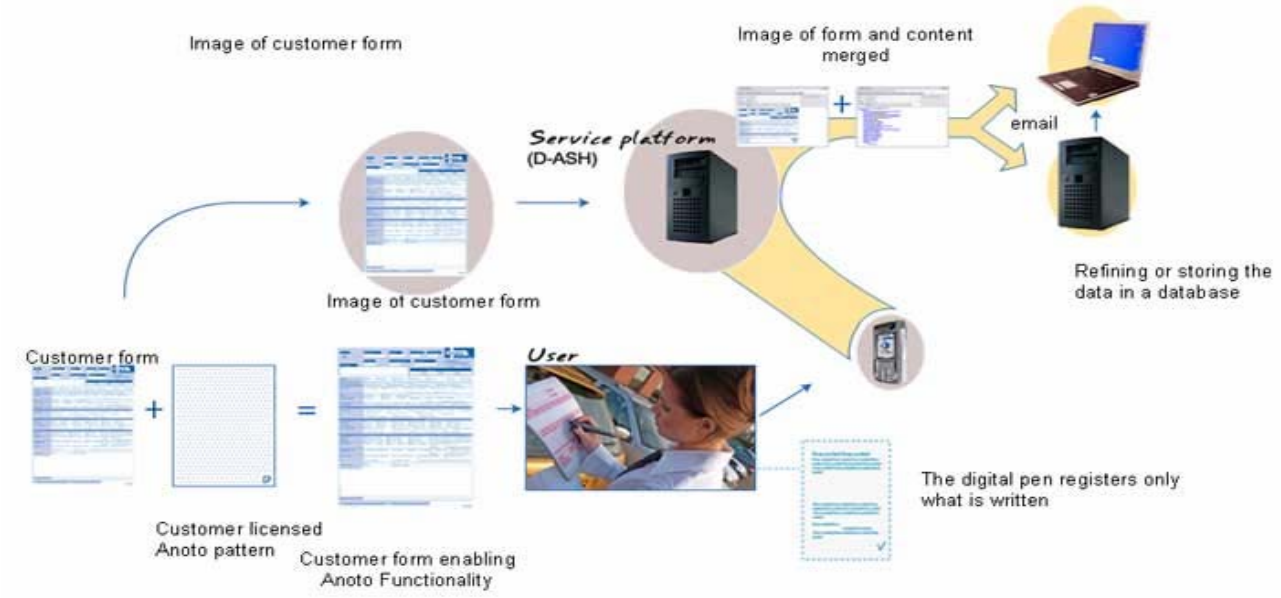
- a technique for detecting locations on a paper substrate that is robust and more reliable than existing solutions using simple electronic sensing;

- a substrate and artwork where the pattern is invisible for practical purposes;
- an information architecture that can support a range of different media;
- a range of interaction styles and a range of different kinds of linking mechanisms;
- interaction styles that are consistent with properties of the media and make apparent the augmented capabilities;
- authoring tools that support professional and bespoke publishers;
- ways of integrating augmented reading with augmented writing and approaches for integrating active paper graphics to provide feedback and additional resources to support augmented reading and writing;
- methods for designing, developing and assessing augmented paper applications considering the needs and requirements of content providers and ‘end users’.

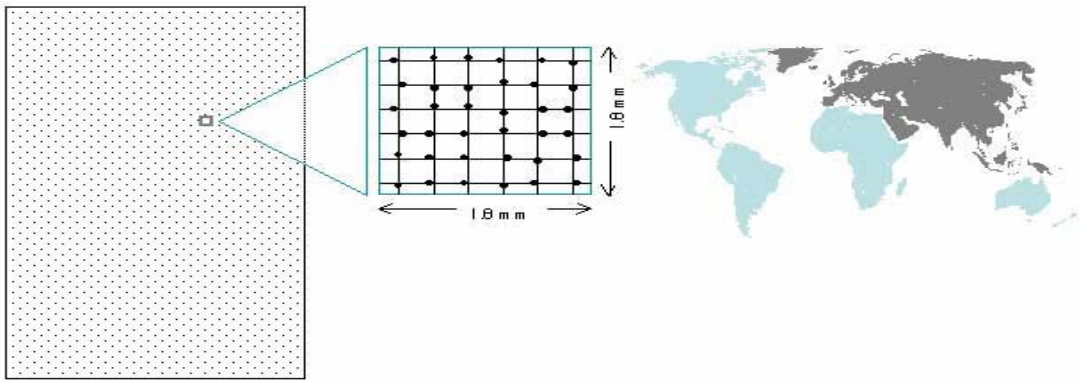
Anoto Technology

One way of detecting positions on paper has been developed by Anoto which forms the basis of commercially available products such as Nokia's Digital Pen, Logitech's Io and Sony Ericsson's Chatpen. These devices capture handwriting, so notes can be sent via e-mail or downloaded to a computer and then converted to text. The Anoto technology relies on an almost invisible pattern of pre-printed dots on the paper and sophisticated electronics built into the pen. Instead of scanning and recognizing single lines of text, the Anoto pen uses a built-in CCD camera to view the infrared-absorbing dots, each of which is slightly misplaced from a square array. Images are recorded and analysed in real time to give up to 100 x-y positions per second, which is fast enough and of sufficient resolution to capture a good representation of all handwriting.
The information is stored as a series of map coordinates. These coordinates correspond to the exact location of the page you’re writing on.





The Anoto pattern is a single unique pattern that if printed out would cover the whole of Europe & Asia.
The pattern can be embedded in any paper document and images such as form layouts can be overprinted.



Once graphical images are received they can then be converted to ASCII text and transmitted in data files most commonly in XML format.
The paperworks project is supporting a variety of applications but this paper looks specifically at an experimental assessment and collaboration application

The Assessment Application- Anoto Technology and Collaborative Assessment

The Original Study

The Technology Education Research Unit (TERU) at Goldsmiths College-London have developed a system of assessment that measures and rewards design innovators working in collaboration.

Collaboration

There is a mass of literature concerning the importance of team-work both in teaching/learning and design settings. But for assessment purposes, there is a pathological fear of using the massive support that it provides to students because of the association with 'cheating'. There is also the difficulty of being able to separate out and award credit for individual contribution to members of the team.
The Technology Education Research Unit were determined to overcome this problem and arrived at a solution involving groups of three students . The grouping was designed explicitly to support and enrich the *individual* work of the team members. It is this individual work that is then assessed.

The Assessment Activity Description

The assessment activity is developed through GCSE Design and Technology, looking at the assessment of generation ideas, development of ideas, and proof of concept.
The booklet which records the students work has two phases of use;

1. during the project exam / design task, by students
2. during assessment, by the assessors

The students work together in groups of three with a head designer and two co-designers, each with their own design booklet.
The assessment activity was developed as a 6-7 hour task: two consecutive mornings of 3 to 3.5 hours. In that time, students start with a task and work through from an initial concept to the development of a prototype solution – a 'proving' model to show that their ideas will work. The whole 7 hours is run by the teacher – following a script that choreographs the activity through a series of sub-tasks - each of which is designed to promote evidence of students' thinking in relation to their ideas.

These 'steps' in the process all operate in designated spaces in a booklet;

1. read the task to the group and establish what is involved
2. explore a series of 'idea-objects' on an 'inspiration table' and in a handling collection designed to promote ideas about how boxes / packages / containers might transform into other forms and functions.
3. put down first ideas in a designated box in the booklet
4. working in groups of 3, students swap their booklets and each team-mate adds ideas to the original
5. team-mates swap again so that each team member has the ideas of the other two members
6. booklets return to their 'owner' and team members discuss the ideas generated
7. the teacher introduces the modelling/resource kit that can be used throughout the 2 mornings
8. students develop their ideas in the booklet – and/or through modelling with the resources
9. students stop to reflect on the *user* of the end product and on the *context* of use, before continuing with development
10. at intervals, students are asked to pause and throw a dice – with questions on each face. The questions focus on procedural understanding e.g. "how would you ideas change if you had to make 100?' and students answer the questions in their booklet

11. photographs are used at approx 1 hr intervals to develop a *visual story line* to illustrate the evolution of models & prototypes
12. at the end of the 1st morning, students – and their team members reflect on the strengths and weaknesses of their evolving ideas
13. the 2nd morning starts with a celebration of the work emerging from day 1. This is based on post-it labels that highlight students' thoughts about the qualities in their ideas
14. further prototype development
15. regular hourly photos and pauses for reflective thought on strengths and weaknesses
16. final team reflections, when (in turn) team members review each others' ideas and progress
17. individually, students then 'fast-forward' their idea illustrating what the product will look like when completely finished and set-up in context
18. students finally review their work from start to finish.

Some Conclusions from the Original Study

Detailed conclusions can be found in the report (reference 1).
An assessment activity was created that was very tightly controlled by the administrator; using the script, the booklet, the handling collection, the modelling resources etc. But learners' reaction to it, reported, that they feel a great sense of freedom in developing 'their own' ideas and making 'their own' products. The procedural framework which was encapsulated by the format of the booklet is the secret to this. It is rich in support systems, creating fertile ground for learners' independent ideas to take root and flourish as well as providing the ideal vehicle for group interaction and interchange.

Team-work

At the end of each activity students were asked to complete an evaluation form which identified all the components of the activity (e.g. the photo story-line). On a Likert scale of 4-1 (very helpful, helpful, unhelpful, very unhelpful) students were asked to identify their reaction to the components. For girls, it is clear that the most popular features of the activity are the *group generation of ideas*, the *photo story-line* and the use of *modelling resources* (all approx 3.5 on the Likert scale*)*, and these three are shortly followed by *the group evaluation of ideas* (Likert 3.4). For boys, six features rank at almost the same level of helpful/very helpful; The *handling collection*, the *group generation of ideas*, the *photo story-line*, the *modelling resources, the booklet space for sketches/notes,* and *helping the group with ideas* (all approx Likert 3.2*)*.

This Collaborative Assessment Study Adapted for use with Anoto Technology

This is work in progress

The students are each given a booklet printed on Anoto paper ,a modelling kit, PDAs, PCs, printer, large screen display and Anoto pens



The Anoto booklets were carefully designed to unfold throughout the activity ensuring students always have sight of the instructions for the sub-task they are currently working on and the work they have just completed.
There are 22 steps which are all represented by a frame to be filled in by the student(s) during the assessment period. All frames contain hints for how to go on with the design process and how to reflect on your own work. There are several different kinds of frames:

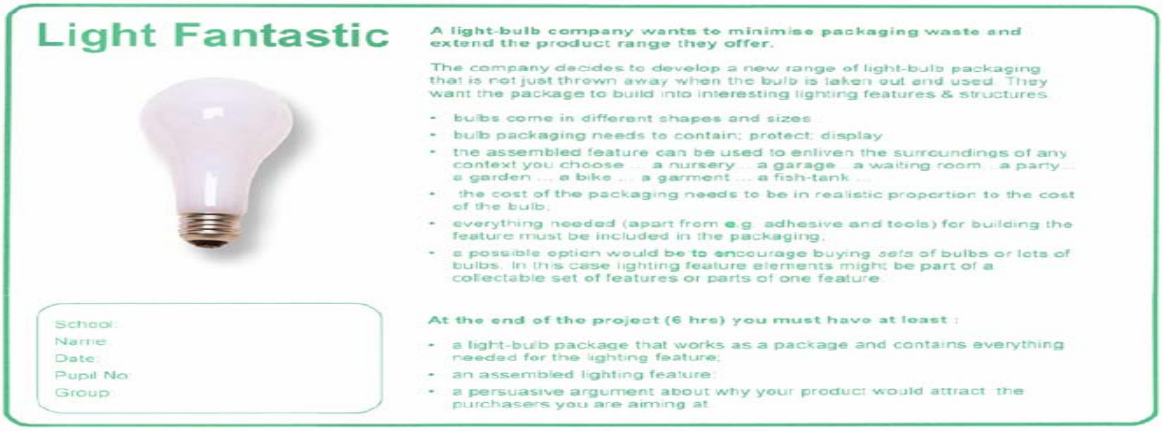
- Text frames with lines to write text on.
- Sketching frames with open space for sketching and writing
- Post-it frames for sticking on Post-it notes
- Picture frames for sticking on digital pictures of models and material made during the project
- Combined frames for ideas, notes, sketches, dimensions, design decisions

There are a series of symbols (tools) which can be used to allow for certain actions to take place or objects to be linked. They appear as buttons on the paper

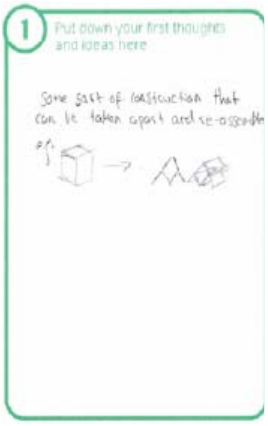
The Anoto pen will recognise the underlying co-ordinates where these symbols are printed and will trigger the appropriate action to take place i.e. replay video, snap to grid to construct a net shape of a 3 dimensional image etc

Scenario 1: Enhanced Design Work

This scenario focuses on how the students use the booklet during the project.
The students have an Anoto booklet, an Anoto pen each, a PDA/computer (possibly projected onto a wall), a digital (video-)camera, a printer and a variety of modelling material for the assignment.
The booklets are folded – only showing the design task
The student's name, group number and group member's names are pre-printed. The groups are formed and registered beforehand.
It is the Anoto paper which is registered to an individual student or a space on the paper. The Anoto pens in this experiment were anonymous and can be used by any student.

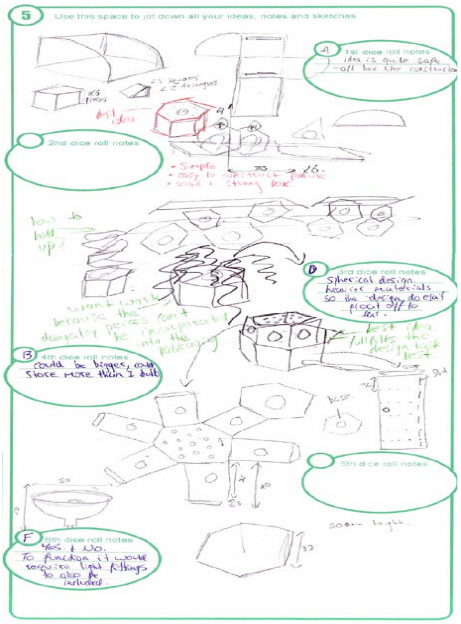


Frames 1,2,3,4



These frames are designed to be worked on by the 2 other partners in the group. The Anoto pattern in this part of the booklet is registered individually to the other two partners
What would the first partner do if this was their own project?
What would the second partner do if this was their own project? At any stage it will be possible to open one or more frames to be displayed on a large screen display. This allows a selection of frames for a group discussion, at any time during the process

Frames 5



Frame 5 extends the collaborative element by allowing the design partners to, not only write down their comments in a particular frame, but to write or sketch their comments and ideas directly onto the sketching surface

Tying identity to the paper, and printing out frame 5 for partners to comment or add sketching ideas, makes it possible for partners to comment and sketch on top of head designer's sketches without disturbing the original drawings

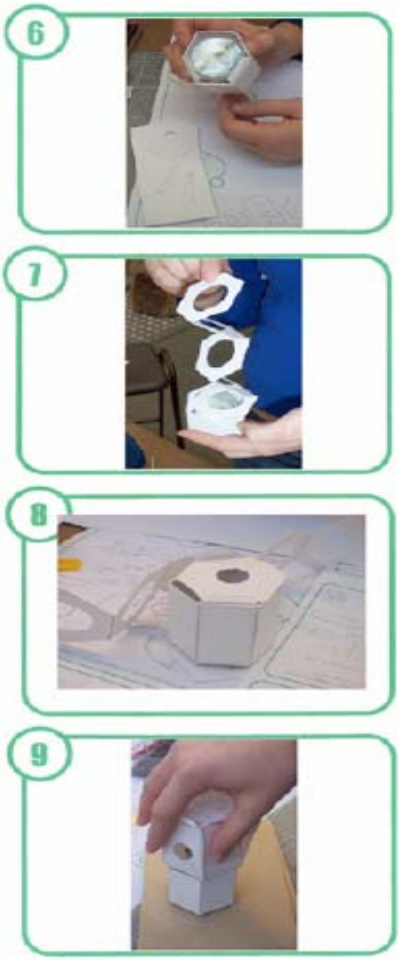
This can be done by having a printing function; printing out the sketch frame for the partners

For the assessor it will be possible to trace who did what, even though pens are used randomly

Print to partner 1

Print to partner 2

Frames 6, 7,8,9



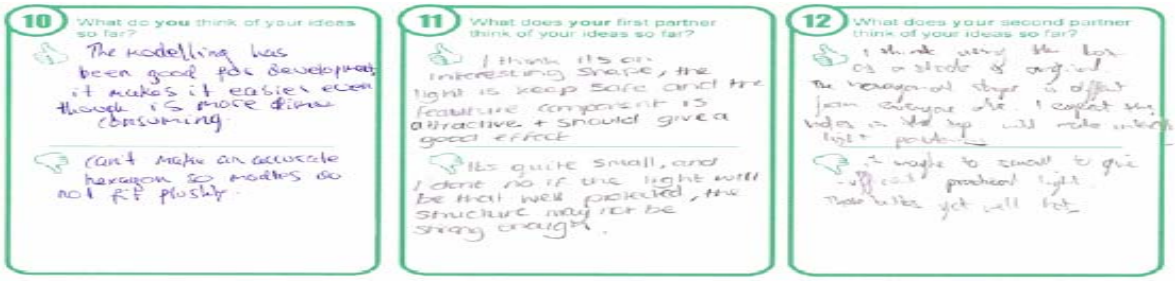
By sketching with the Anoto pen communicating with a computer the hand-drawn sketches in frame 5 can be augmented by ‘snapping’ the shapes onto grids. Software on the PC facilitates this process. This could be used to build the physical model. In this example the 6-sided polygon can be drawn with precise sides and angles, and in the preferred dimensions. It will be easy to change the size. Linking can be made to inspirational ideas like shapes from nature, industrial products, and the human body by simply adding a link to a website or linking to a document on the computer. The sketches can be displayed on the wall in order to enhance collaborative design decisions. The model can be transformed to a 3D model. After refining the model from paper sketch to 3D in the computer, the computer model can be ‘unfolded’ and printed on cardboard in order to make a cardboard prototype.

Snap to grid

In frame 6-9 the first cardboard based prototype is documented by picture-links to short videos showing different aspects of the design concept.
Photo series of early model (day 1)

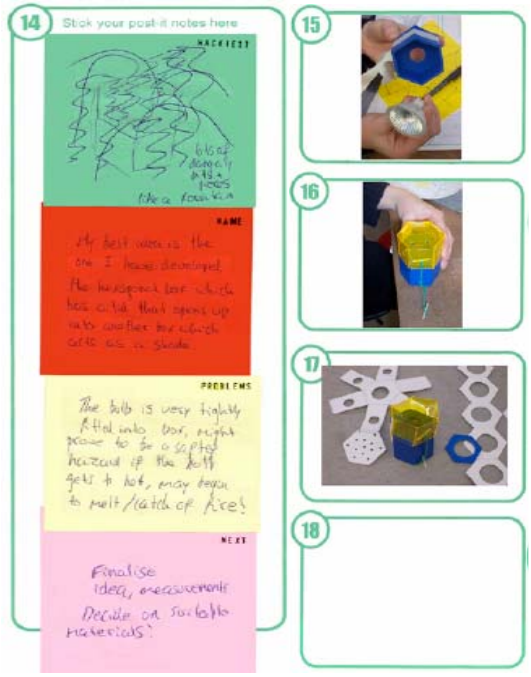
- The photos can be video clips
- It is possible to initiate the video clip by touching the play button with Anoto pen

Frames 10,11,12



Frame 10-12 lets the students discuss aspects of the design concept so far, by writing down positive and negative opinions from each of the three students. i.e. What do you think of your ideas so far? (10), What does your first partner think of your ideas? (11), What does your second partner think of your ideas? (12)
One or more particular frames can be displayed on a large screen, for group discussions

Frames 14 -22



Frame 14 is for sticking on Post-it notes with keywords summarizing the design sessions of the day among all groups. This is done at the end of day 1 on a large wall display and 'recorded' by the Anoto-pen. Each Post-it sticker contains links to the individual projects.
Frames 15 -18 contains information on how a model is built in materials which are representative of the final design solution. This part is documented in frame 15-17 (18) by picture links to short videos. The pictures contain links from different parts of the prototype to material specifications, samples of the material and other designs made in the same material.
Photos from day two. The photos can be video clips which can be activated by touching the play button with Anoto pen.
Frames 19 and 20 contains partners 1 and 2's thoughts on the design ideas so far.
Frame 21 contains the 'head' designer's own thoughts on the work.
Frame 22 contains possible obstacles for the future

Future Work; Scenario 2: Enhanced Assessment

This is the next phase of the study and is already in progress, but at time of writing this paper specific results are as yet to be collected and analysed. This scenario focuses on how the examiner can be supported in the assessment of the work.
The system is being designed to allow the assessor to track:

- design decisions
- distribution of work and decisions among students
- external resources (links made during the design work)
- speed of work
- process
- interaction and collaboration amongst the students

There are several work parcels supporting the assessment process. The first will look at compiling & collating files for assessment. Files can be layered to allow synchronous and asynchronous work to be viewed. It will analyse the physical process of assessor interaction with a variety of digital information

i.e. zooming in and out and multiple screen viewing. The second will look at the assessment tools with the development of the assessment tool box i.e. instant access to assessment criteria, archive materials, benchmark standards and exemplar matches.

Conclusion

In a digital world, paper and pen have an important but potentially new existence.
Studies have discovered over and over again one remarkable fact; despite the popularity of new technologies, paper and the use of the pen remains a critical feature of work based activities and collaboration.
The challenge for both digital and paper/pen is to interrelate and bridge this apparent divide. This is a challenge that has only recently been recognised.
In this study there is a recognition that paper provides the vehicle, space and medium for collaboration between individuals working as part of a group. The study developed a paper based procedural framework in the form of an assessment booklet. This paper based framework sets out a structure for collaboration, group interaction and interchange.
The Anoto technology harnesses the successful features of this collaborative paper framework and in addition provides a variety of tools which enhance the assessment of group and individual work. These tools include a facility to a replay the students work as he or she goes through the development process. As the student inputs information in the form of drawings, writing and voice files, photographs and videos, this information is time stamped. This allows each piece of information to be tracked. It also gives a greater insight into the students thought process and evolving ideas. A variety of conclusions can be drawn from this insight i.e. a greater understanding of the interchange of ideas within the group, looking at how one persons idea has influenced another students thinking. Also in other assessment scenarios it may be useful in the diagnosis of conceptual misunderstandings. The learner can then be guided towards an appreciation of other ways of viewing, understanding and or working out that concept.
The third phase of this study will concentrate on how assessments can be made. To develop an understanding of what it is we are assessing in this new interactive and collaborative world. To look at the perceived benefits and efficiency that may be offered by connecting digital and paper based activities.
The study so far has provided a brief glimpse of what may be awaiting. This work is in its infancy but is providing the foundation for a rich and varied source of future research, analysis and development in a newly emerging field.

References

1. **Assessing Design Innovation-** A research & development project for the Department for Education & Skills (DfES)and the Qualifications and Curriculum Authority (QCA)by the Technology Education Research Unit Goldsmiths College: University of London December 2004, Richard Kimbell, Soo Miller, Jenny Bain, Ruth Wright, Tony Wheeler, Kay Stables
2. **Tasks-in-interaction: paper and screen based documentation in collaborative activity** in Proceedings of CSCW ‘92, Toronto, Canada. Paul Luff, Christian Heath and David Greatbatch
3. **Documents and Professional Practice: 'bad' organisational reasons for 'good' clinical records** Proceedings of the Conference on Computer Supported Cooperative Work, Boston: ACM Press 1996. Christian Heath and Paul Luff, Centre for Work, Interaction and Technology, School of Social Sciences. University of Nottingham
4. **Mobility in Collaboration** Paul Luff, WIT Research Group, King’s College London, Christian Heath, WIT Research Group, King’s College, London, in Proceedings of CSCW'98. (Seattle, WA) ACM Press,1998.

DEVELOPING A ROADMAP FOR E-ASSESSMENT: WHICH WAY NOW?

Denise Whitelock and Andrew Brasher

Developing a Roadmap for e-Assessment: Which Way Now?

Denise Whitelock and Andrew Brasher
The Open University
Walton Hall
Milton Keynes
MK7 6AA
d.m.whitelock@open.ac.uk

Abstract

e-Assessment is of strategic importance to the UK since it forms an integral part of the e-learning movement which is a major global growth industry. This paper reports results from a project commissioned by JISC which set out to develop a Roadmap for e-assessment.

This methodological approach was drawn from a range of 'roadmap' methodologies collected by Glenn and Gordon (2003). It facilitated the identification of the enabling factors and barriers to the use e-assessment through the construction of a survey which probed a number of experts opinions.

The analysis of the various sources suggest that in England and Wales it is policy pressure which is a main driver and it is affecting more of the FE sector than the HE sector. In the HE sector institutions have more control over the rate and uptake of e-assessment as they award their own degrees. However, there is a recognition in HE that with larger classes and less tutorial time, tutors can keep track of their students' progress through e-assessment systems. They can adjust their lectures accordingly after they have picked up the misconceptions of a cohort through e-assessment feedback. At a personal level teachers/enthusiasts are addressing pedagogical problems through e-assessment.

The barriers identified at a superinstitutional level, for example the . DfES, funding bodies, and examining bodies, are that of regulation, confidentiality and testing of these systems before they go across the UK. While the main drivers at a superinstitutional level are to move towards a new generation of learners engaed in self-reflection who will be able to identify their own learning needs. One of the major drivers for institutions to adopt e-assessment practices is that of student retention. HE and FE also see benefits with respect to attendance and achievement. This paper outlines the methods used and describes key barriers which will have to be overcome if e-Assesment is to be effectively deployed across UK HE and FE sectors. .

Introduction

This project, based at the Open University, set out to review current policies and initiatives relating to e-Assessment across the UK, as documented by the funding councils, examination boards and accrediting bodies. Strategic priorities, projects and research activities were identified to assist with the development of recommendations for future coherent development in this field. This was achieved through not only suggesting ways to implement such policy documents as the DfES Harnessing Technology (2005) report, but also by adding value to the teaching and learning sector, through the advice of known experts gained during the development of the roadmap. This outcome was progressed through a modus operandi which selected a number of facets from a range of roadmap methodologies collected by Glenn and Gordon (2003).

e-Assessment is defined in its broadest sense, where information technology is used for any assessment-related activity. e-Assessment can be used to assess cognitive and practical abilities. Cognitive abilities are assessed using e-testing software, while practical abilities are assessed using e-portfolios or simulation software (Wikipedia) http://en.wikipedia.org/wiki/e-assessment.

This paper summarises the factors influencing the methods adopted, describes the methods, and gives an overview of some of the key findings in terms of barriers that have been identified.

Roadmapping Practice

In general the aim of a technology roadmap is to provide a consensus view or vision of the future landscape available to decision makers. The roadmapping process should provide a way to identify, evaluate, and select strategic alternatives that can be used to achieve a desired science and technology objective (Kostoff and Schaller 2001). In the case of this roadmap, the science and technology objective can be summarised as ‘effective implementation of e-Assessment within the post-16 and higher education sectors’. This roadmap seeks to present a vision of the future landscape that will help organisations and individuals in the post-16 and higher education sectors to make decisions about their future plans with respect to e-Assessment.

A chapter on science and technology roadmapping (Gordon, 2003) in an extensive survey of futures research methodologies (Glenn & Gordon, 2003) states:

“Since a roadmap is a diagram of interconnected nodes, it is necessary to consider what a node and the interconnections – that is the lines connecting the nodes – represent.

A node is a milestone on the road being mapped. It can be an element quantitatively determined (e.g. a document which is cited, a patent which is represented by other patents as a precursor) or subjectively defined (e.g. a future technology at some level of performance). When the node is

quantitative, the definition can be “looked up” in some data base; when it is qualitative, usually the node is determined by expert opinion.”

Roadmaps are used for both retrospective and prospective studies in time, the link vectors can assume forward and backward directions in time. Construction of a roadmap, thus, requires identifying the nodes, specifying the node attributes, connecting the nodes with links, and specifying the link attributes.

There can be many approaches to developing such a roadmap. However surveys of approaches (e.g. Gordon, 2003; Kostoff & Schaller, 2001) indicate that what is required in for considering future directions is a *prospective roadmap* i.e. a map to help find out where we are going, as opposed to a *retrospective roadmap* which is intended to tell how we got to our present position. Kostoff and Schaller identify two extremes of prospective roadmap

Requirements-pull roadmaps (which start with desired end products and fill in the remainder of the roadmap to identify the R&D necessary to arrive at these products)
and

Technology-push roadmaps (which start with existing research projects, and fill in the remainder of the roadmap to identify the diversity of capabilities to which this research could lead).

For this project, we required a method that takes account of both requirements-pull and technology-push because we recognise that the development of this roadmap must consider political, pedagogical and business drivers for e-Assessment technology in addition to R&D showing how technology can be appropriated and used to support assessment. Factors influencing the choice of methods included

- Duration and budget of the project
- Availability of expertise outside the project team
- Reports and policies specified to be relevant by JISC.

Methodology

The project divided into three main stages, as illustrated in Figure 1. The methods used within each stage are described in the next three sections.

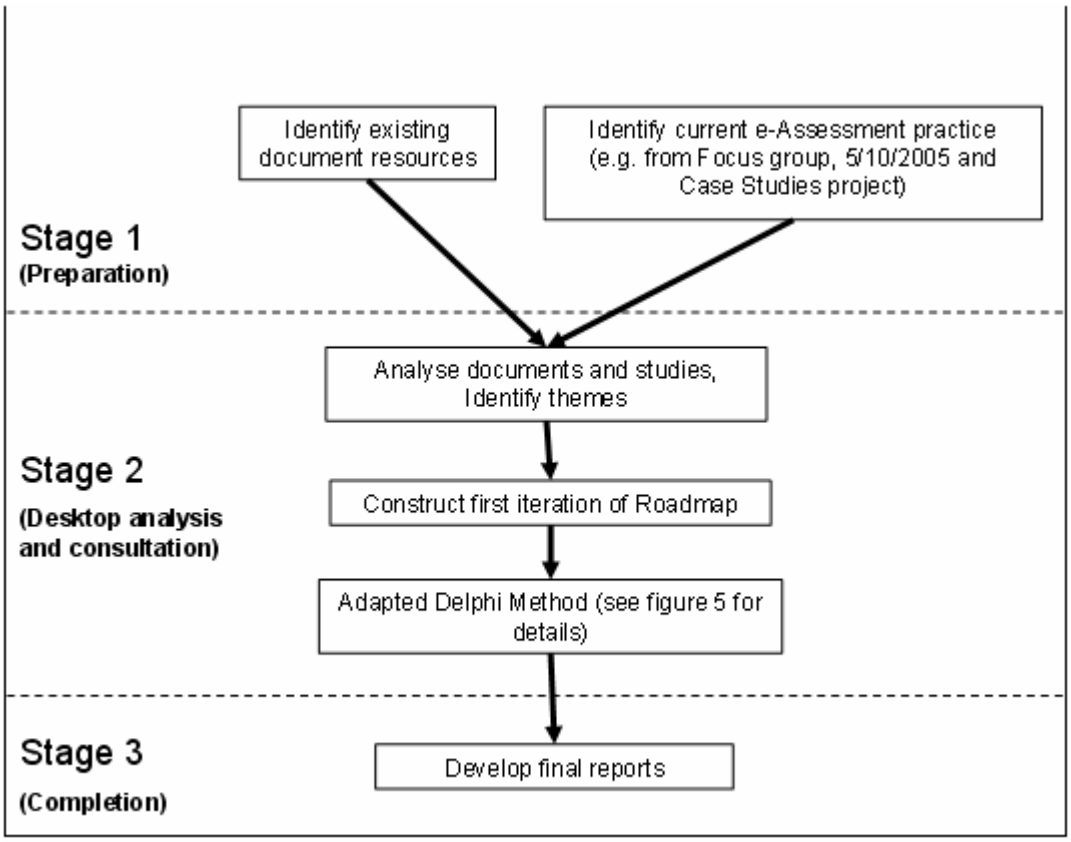


Figure 1: Graphical representation of the stages in the project

Stage 1: Preparation Phase

Stage 1, the preparation phase set out to achieve two goals.

1. Identifying key documents
   In consultation with JISC a number of UK organisations considered to be important players with respect to assessment in general and e-Assessment in particular were identified. Policy and other documents which described the plans and policies of these organisations with respect to e-Assessment were identified and obtained, together with published academic papers, about the role of standards, development of automated marking systems and pedagogical drivers to adopt e-assessment. (HEFCE, 2005; 14-19 Education and Skills White Paper, 2005; QCA Blueprint for e-Assessment, 2004; The development of e-Assessment 2004-2-14 report, 2005; SQA Guidelines on e-Assessment for Schools, 2005; DfES Harnessing Technology, 2005)

2. Identifying current e-assessment practice
   Sources of information about the current state of the art in practice, and the future plans of leaders in the field, were identified. This database of current e-Assessment practice was complied through close co-operation with the JISC e-Assessment Case Studies project (Case studies of effective and innovative practice in the area of e-assessment http://kn.open.ac.uk/public/index.cfm?wpid=4927) which also involved members of the Open University's roadmap team.

Stage 2: Desktop Analysis and Consultation

An analysis of the key documents and the database of current e-Assessment practice identified in Stage 1 was carried out, to identify the strategic issues and challenges and benefits of e-Assessment together with the institutional, operational and pedagogic enablers and barriers to the effective use of e-Assessment. This analysis led to the development of a framework for constructing a first iteration of the roadmap.

Roadmap Framework

The outcome of the literature review and the analysis of the database of current e-Assessment practice inspired the framework shown in figure 1. This framework consists of two axes, 'Status' and 'Scope' each consisting of three cells. The cells along the 'Status' axis (i.e. 'Vision', 'Barriers' and 'State of the Art') represent the current status ('State of the Art'), a vision of a desirable future status ('Vision'), and barriers which will need to be crossed to reach this desirable vision from the current status.

The cells along the 'Scope' axis (i.e. 'Superorganisational', 'Organisational' and 'Personal') represent the organisational scope to which the roadmap nodes within the cells of the map will apply. 'Personal' scope means e.g. the scope of individual academics or students. 'Organisational' scope means e.g. the scope of academic or commercial organisations involved in e-Assessment activities. 'Superorganisational' scope means e.g. the scope of those bodies which represent the interests of more than one organisation. Examples of 'Superorganisations' include government departments (e.g. DfES), funding bodies, and examining bodies. In Figure 2 the text in each cell gives an example of the nature of the nodes which will occur in each cell.

Figure 3 extends the framework shown in Figure 2 to include an indication of the linkages which this form of map will show. These linkages are representative of strategies and facilitators that will help overcome the barriers and facilitate organisation, superorganisations and people change their status from their current state to the desirable vision.

Scope / Status	Superorganisational	Organisational	Personal
Vision	What the superorganisational policy documents are aiming for in terms of e-assessment.	What organisations are aiming for in terms of e-assessment (make use of case studies?)	What individuals are aiming for in terms of e-assessment? Teaching staff, students, researchers, others?
Barriers	Misalignments between policies of various superinstitutions.	Misalignments between superorganisational policies and capabilities and/or policies of organisations.	Barriers which prevent (or reduce effectiveness of) individuals becoming involved in e-assessment.
State of the Art	Current situation in 2006	Current situation in 2006	Current situation in 2006

Figure 2. Roadmap framework

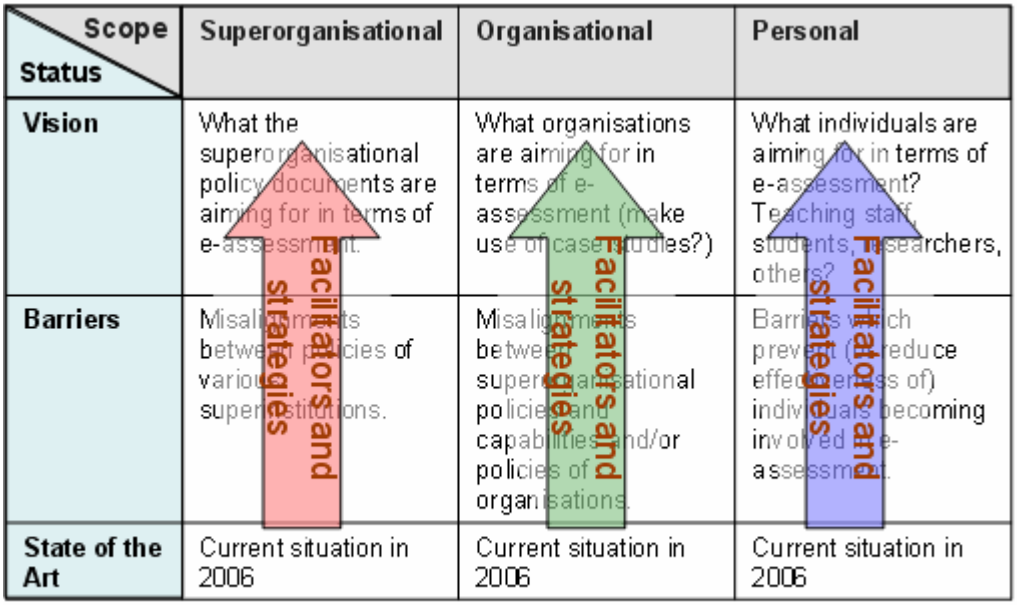


Figure 3. Roadmap framework showing linkages between nodes and cells

The purpose of Figure 4 is to clarify the framework by describing the meaning of facilitators, strategies and barriers within each scope category.

	Superorganisational	Organisational	Personal
Vision	What the superorganisational policy documents are aiming for in terms of e-assessment.	What institutions are aiming for in terms of e-assessment (make use of case studies?)	What individuals are aiming for in terms of e-assessment? Teaching staff, students, researchers, others?
Barriers	Misalignments between policies of various superorganisations.	Misalignments between superorganisational policies and capabilities and/or policies of institutions.	Barriers which prevent (or reduce effectiveness of) individuals becoming involved in e-assessment.
Facilitators	Alignments between policies of superorganisations. E.g. systems not directly related to e-assessment, but the existence of which facilitates e-assessment.	Alignments between superorganisational policies and capabilities and/or policies of institutions.	Actions and processes which promote individuals involvement and/or gains from e-assessment.
Strategies	Suggestions about how to move through the barriers towards the vision.	Suggestions about how to move through the barriers towards the vision.	Suggestions about how to move through the barriers towards the vision.

Figure 4. Table intended to clarify the Roadmap framework

Survey

The main test instrument was a survey sent to a group of experts, comprising of academics, commercial producers and personnel working for Government agencies such as SQA, Becta etc. This survey was an adaptation of the Delphi Method (Gordon, 2003) which makes use of a panel of experts and aims to build consensus over a range of issues. 40/50 returned the survey, a good response rate.

The survey (was designed after the literature review had been completed and key issues identified. Although termed 'Survey' it was more of an electronic consultation as the experts were asked to give their opinions and to write free text responses for 13/16 questions.

The survey probed experts' opinions on the following issues:

(a) The timings of policy implementation i.e. their realisation in HE and FE (2009 deadline by QCA, not so in Scotland)

(b) The way in which e-assessment can make a significant contribution to cutting the burden of quality of assessment

(c) Ways in which e-assessment will make a significant contribution to improving quality of e-assessment

(d) The implications for the vision set by the policy documents (some maybe unforeseen)

(e) Visions for the future

The project's Steering Group and Advisory Group formed the basis of the group of experts for this consultation phase of the project. Please find current list of participants in Appendix 4. The Delphi method was used to test the project team's initial conception of the roadmaps, and to identify factors that may have been omitted.

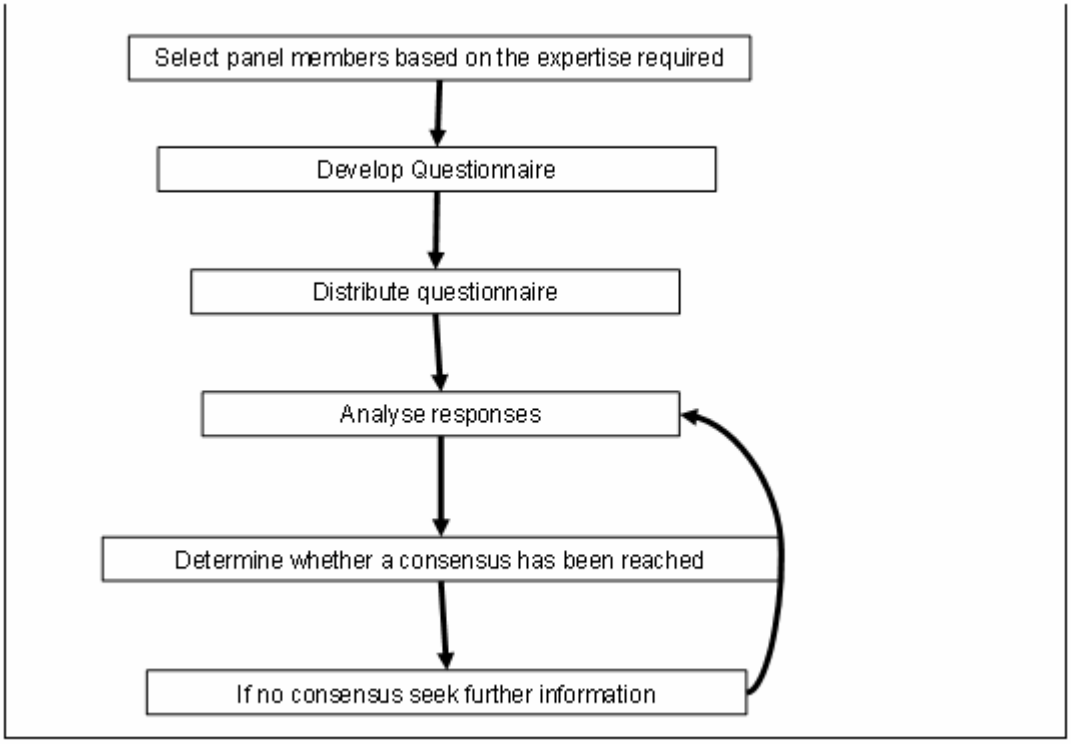


Figure 5. Illustration showing how the Delphi Method was used

Stage 3: Completion phase

Analysis of the results from the Delphi Survey and the literature review enabled the production of a roadmap that illustrates the planning of future e-assessment developments and strategic drivers and initiatives relating to e-assessment.

This includes a visual representation of the roadmap which was produced and implemented using a graphic design tool

Results: What do the experts think? Electronic Consultation (Survey) Findings

The purpose of the electronic consultation was to clarify whether the visions and directives issued by the policy makers in the UK were viewed, by a group of experts in the field, as realistic and matched current progress in the HE and FE sectors. The experts' opinion was also prompted about whether there were any unforeseen or undesirable consequences to the vision promulgated by the Superorganisations. Our group of experts were also asked to comment upon their own visions of the future and to articulate any barriers that they envisaged would deter or prevent educational institutions from piloting e-assessment applications. This section of the paper reviews the following issues:

(a) The timings of policy implementation i.e. their realisation in HE and FE (2009 deadline by QCA,)

(b) The way in which e-assessment can make a significant contribution to cutting the burden of quality of assessment

(c) Ways in which e-assessment will make a significant contribution to improving quality of e-assessment

(d) The implications for the vision set by the policy documents (some maybe unforeseen)

Synopsis of Findings

(a) Predicted timings of e-assessment

Most experts expect e-assessment to make a significant contribution to both the quality and usage of assessment in general by 2010. They also believed that ICT will be commonly accepted into all aspects of the student experience within 2/4 years. Students too will be able to access information, tutor support, expertise and guidance online and will be able to communicate with each other wherever they are within 2/4 years. The consensus view also contained a belief that tutors will have tools for course design and will be able to give better feedback electronically to students again within the next 2/4 years. Therefore, the timings to implement HEFCE strategies with respect to the above-mentioned technologies are considered to be imminent and to match HEFCE's predictions. The recent calls for software development by JISC also support this notion.

(b) The way in which e-assessment can make a significant contribution to cutting the burden of quality of assessment

The experts believed that the introduction of technological change can facilitate reflection upon our practice and encourage a significant revision of current e-assessment customs. They acknowledge that the construction of good e-assessment questions requires change but in the long run good e-assessment would create efficiencies in results processing and transparency

of grading. It will produce faster feedback and that it's main effect will be seen in formative assessment practice which will encourage the students to take control of their own learning.

(c) Ways in which e-assessment will make a significant contribution to improving quality of e-assessment

The experts agreed that regular feedback to students in both formative and summative assessments will particularly assist those who regularly under-perform. There will be more evident changes in the vocational sphere but a wider range of curriculum will be tested by e-assessment. This will be because more realistic assessment such as problem solving scenarios will be offered to students.

(d) The implications for the vision set by the policy documents (some may be unforeseen)

Experts agreed that the over-use of results from on-demand testing does not always increase grades and can lead to a lack of confidence in standards by the general public. They also suggested that if the vision for on-demand testing, as set out by the Government, is implemented then this will mean e-assessment sites will be open 24 hours a day. One of the unforeseen implications for this policy could be that parents will over-pressurise children to take exams too early. Also more students will probably study university courses while still at school.

Visions of e-Assessment for 2014

The experts have a coherent vision that e-assessment can assist learning and expect more formative e-assessment to be available to students. The effect of this development will be to encourage students to check their understanding of a given topic more frequently. The experts do not expect unassisted practice alone will aid learning but the quality of feedback given to the students will encourage reflection and enhance learning.

Delivery of e-assessment 2014

Superorganisational

The experts agreed that on-demand testing will be available for AS and A Levels. They were less confident that this would be the case throughout the HE sector. They suggested that large-scale testing sites would be available as now organised for the theory driving test. These testing sites could be located in schools, colleges, universities and possibly supermarkets. e-Assessment will be prevalent from primary school through to university and other institutes of higher education. However high stakes assessments will still be available in traditional forms such as the final examinations taken at university level. They acknowledge there will be set backs which will reduce confidence in e-assessment and progress could be slower than expected.

Organisational

The experts believe that e-portfolios will play a large role in the assessment of courses delivered both in FE and HE institutions. Formative and self assessment together with e-portfolios will make up a core of assessment tools. There will be a change in competence measurement as this will occur at random intervals rather than as a series of discreet controlled events.

Personal

Some of the personal visions revealed some blue skies thinking where some of the experts predicted that e-portfolios could be exchanged as microchips in a business card, that e-assessment will replace everything except practical examinations and it will be integrated seamlessly into day to day learning and work environments.

Visions for Research and Development

Superorganisational

The experts suggested that a set of guidelines will be available to ensure the quality, accessibility, reliability and security of all e-assessment tasks. They did not believe that research and development of into standards should dominate the research agenda or slow down development of systems driven by pedagogical need.

Organisational

This group proposed that the development of quality training programmes for teachers, developers and invigilators will be delivered electronically. There will be a set of excellent tasks available to assess group work electronically. There will also be peer e-assessment together with adaptive systems that respond to students' misconceptions during formative assessment tasks. They also believed that the use of virtual reality technology will increase the authenticity of certain assessments. More unusual uses of technology will also be prevalent to assist with learning that is more personal.

What are the Barriers to these Visions?

The expert group contributed to a variety of issues which may hinder development and adoption of e-assessment. These are grouped into those that will:

- affect the widespread adoption of proven systems (i.e. systems which have been proved to work in pilot studies)
  and others which will
- hinder the initial research and development of e-assessment systems.

Barriers to Adoption: Superorganisational

The superorganisational barriers identified by the experts were concerned with a lack of customer confidence in the awards accredited using e-assessment systems. Problems associated with e-assessment pilots will reduce confidence and also where people believe that current systems are doing a good job and therefore they do not need to be changed. To move e-assessment forward there needs to be enough resources available and appropriate technical infrastructure should be in place. There also needs to be a commitment at a technical level to achieve interoperability of systems across institutions. Another barrier is the lack of sharing of best practice among institutions.

Barriers to Adoption: Organisational

The experts suggested that more institutional "buy in" is required and hence a culture shift is needed to change both the planning and business processes to fit new assessment practices. A lack of staff skills and expertise was noted and so training needs to be put in place. One of the major barriers recognised was the time required to develop good e-assessment tasks. Staff therefore need to be given time and recognition to carry out this work.

Barriers to Adoption: Personal

At this level experts recognised the work pressures on academic staff to produce good e-assessments, that there were training needs that had to be addressed, together with learner attitudes. The latter need to have confidence in the security and marking of the e-assessment assignments. Learner scepticism that e-assessment can be a valid way of examining key skills to post degree level needs to be addressed.

Barriers to Research Development and Piloting: Superorganisational

The experts mentioned the following two major barriers which included customer attitude and lack of public confidence in e-assessment. They also mentioned the lack of integration of institutional and Government policies to ensure that the key criteria of quality, accessibility, reliability and security are evaluated in future pilot activities.

Barriers to Research Development and Piloting: Organisational

A lack of funding to encourage institutions to engage in pilot and/or research and development activities was mentioned. A lack of resources and suitable infrastructure to pilot e-assessments was also high on the experts' list of barriers.

Barriers to Research Development and Piloting: Personal

The barriers here fell into three major categories. Those of staff attitude where a lack of encouragement to individual practitioners was mentioned with respect to limited funding and time and recognition to continue working in this

area as opposed to personal research time which has more RAE status. A lack of infrastructure and also ICT skills in the student population to pilot projects was also recorded.

Summary

Findings suggest that in England and Wales it is policy pressure which is a main driver and is affecting more of the FE sector than HE sector. HE has more control over the rate and uptake of e-assessment in their institutions as they award their own degrees. However, there is a recognition in HE that with larger classes and less tutorial time, tutors can keep track of their students' progress through e-assessment systems. They can adjust their lectures accordingly after they have picked up the misconceptions of a cohort through e-assessment feedback. At a personal level teachers/enthusiasts are addressing pedagogical problems through e-assessment.

The barriers identified at a superorganisational level are that of regulation, confidentiality and testing of these systems before they go across the UK. Also there is more reliance than expected on the private sector and small commercial businesses to achieve the vision. Providing e-assessment systems is expensive and some institutions have invested heavily in particular VLEs. They in turn have their own 'e-assessment systems'. In practice some of these are little more than quizzes and do not meet the aspirations of institutions who want to pursue interactive assessment systems which also provide instant feedback to students.

Teachers themselves are not convinced that e-assessment can test enough learning outcomes. They are also concerned about plagiarism and require more training to use and develop questions.

The main drivers at a superorganisational level are to move towards a new generation of learners engaged in self-reflection who will be able to identify their own learning needs. One of the major drivers for institutions to adopt e-assessment practices is that of student retention. HE and FE also see benefits with respect to attendance and achievement. Accreditation can also be tracked through e-assessment systems.

Tutors want to use e-assessment especially formative e-assessment as diagnostic tools to understand how their students are learning especially in larger groups. They can then adjust their teaching accordingly and we have noted changes in pedagogical practice with the introduction of e-assessment (case studies project). There is a recognition at University level that more research funding is needed for e-assessment especially in the area of text recognition and automated feedback. In a sense more joined up thinking is needed at superorganisational level where there should be more of a push to ensure technical standards are in place and that there is a code of practice developed with guidelines as well as industry standards. Institutions are developing but need to make more explicit their e-assessment policies and invest in staff training. Individual champions and teachers would like more

recognition of their work by the VLEs and other commercial software production houses because they are developing systems that address their own particular student needs. They would like these rolled out instead of trying to match their needs to a generic system. In one sense pedagogical needs are hampered by straight jacket software systems and this is where JISC funding can support local champions to build and then develop open-source products. This seed funding in turn fosters take up and further development by other institutions of these pedagogically pertinent systems.

All experts from this group believed in e-assessment becoming integral to teaching and learning in 2014. Although some scepticism about the timing of progress was evident the feeling from this group can be summarised by one member who said:

“I do share the vision expressed in the DfES report – I have done so all my working life really and despite the frequent experience of seeing hopes for the greater use of e-learning deferred, I really do think that ICT in society has now crossed a rubicon and rapid progress is inevitable.”

References

14-19 Education and Skills White Paper. Summary: 14-19 Education and Skills Presented to Parliament by the Secretary of State for Education and Skills by Command of Her Majesty February 2005. Last updated 10 October 2005. Author: Ruth Kelly

DfES (2005) 'Harnessing Technology: Transforming learning and children's services'. Summary: this document was downloaded from http://www.dfes.gov.uk/publications/e-strategy/. The page states "On 15 March 2005, the Department for Education and Skills published the e-Strategy 'Harnessing Technology: … Last updated: 26 September 2005, Author: DfES

Glenn, J.C. and Gordon, T.J. (2003) Futures research methodology-Version 2.0 – CD-ROM. American Council for the United Nations University

Gordon, T.J. (2003) The Delphi Method. In Futures research methodology Version 2.0 CD-ROM. Ed. J. C. Glenn and T.J. Gordon, American Council for the United Nations University, 3.

HEFCE (2005) Strategy for e-learning. Summary: This document sets out our strategy and implementation plan for supporting higher education institutions to develop and embed e-learning over the next 10 years. It reflects responses to the consultation … Last updated: 26 September 2005, Author: HEFCE http://www.hefce.ac.uk/pubs/hefce/2005/05_12/05_12.doc

QCA Blueprint for e-assessment (2004)http://www.qca.org.uk/2586_6997.html

SQA Guidelines on e-assessment for Schools (2005) http://www.sqa.org.uk/sqa/sqa_nu_display.jsp;jsessionid=0CFB9D4DEAAEE06F5F55A97BE494A421?pContentID=6271&p_applic=CCC&p_service=Content.show&

The Exam on Demand Assessment Advisory Group (2005) The development of e-assessment 2004 to 2014 http://www.examondemand.co.uk/Papers/EXoD%20Report%20final%20no2.pdf

IDENTIFYING INNOVATIVE AND EFFECTIVE PRACTICE IN E-ASSESSMENT: FINDINGS FROM SEVENTEEN UK CASE STUDIES

Denise Whitelock, Don Mackenzie, Christine Whitehouse, Cornelia Ruedel and Simon Rae

Identifying Innovative and Effective Practice in E-assessment: Findings from Seventeen UK Case Studies

Denise Whitelock[1] , Don Mackenzie[2], Christine Whitehouse[2], Cornelia Ruedel[2] and Simon Rae[1]
[1]The Open University, [2]University of Derby
d.m.whitelock@open.ac.uk

Abstract

The aim of this JISC funded project was to extend the understanding of what e-assessment meant to users and producers in the HE and FE sectors. A case study methodology was employed to identify and report upon best and current practice within this field of inquiry. This approach facilitated the identification of both the enabling factors and barriers associated with e-assessment.

The variety of applications of e-assessment studied and their innovation and general effectiveness indicate the potential of e-assessment to significantly enhance the learning environment and the outcomes for students, in a wide range of disciplines and applications.

Introduction

The implementation of electronic examinations is being investigated at school, Further Education and university levels throughout the UK. The 14-19 Education & Skills White Paper presented to Parliament by The Secretary of State for Education and Skills in February 2005 states: “In the medium term we expect e-assessment to make a significant contribution to cutting the assessment burden and to improving the quality and usage of assessment”. This research which set out to identify current innovative and effective practice also starts to investigate whether e-assessment can match the claims made by the DfES.

A case study methodology (Gomm et al, 2000) was adopted in order to create a narrative framework within which the barriers and facilitators of e-assessment practice could be contrasted. This research approach was preferred since it not only offered an insight into the design of the overall study as shown in Figure 1 below, but also enabled ‘how’ and ‘why’ questions to be explored in depth during the course of the interviews (see Yin, 2003).

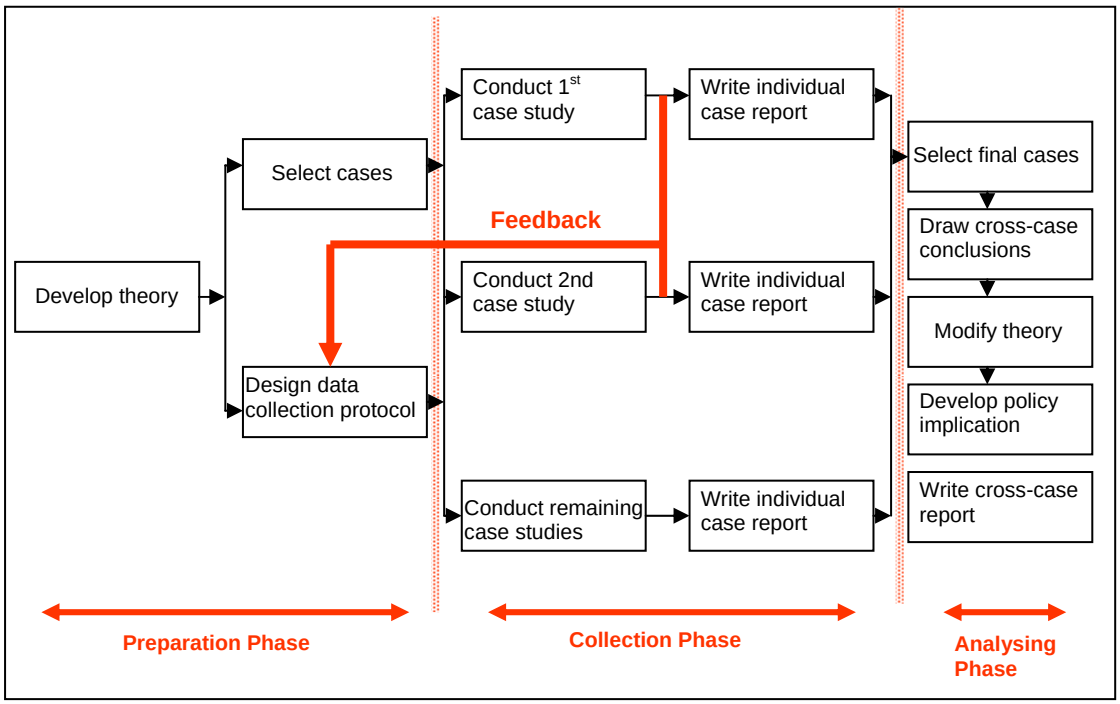


Figure 1: Case study methodology (Yin, Case Study Research, p.49)

The case studies spanned the HE and FE educational sectors in England, Northern Ireland, Scotland and Wales. Three further cases were incorporated from other sectors, including a contribution from the most recent British Citizenship Test, a continuing professional development application for nurses at Chesterfield Royal Hospital and a study of the Cambridge University's on-line admissions test, developed by Cambridge Assessment. Key personnel from twenty different sites were interviewed. These included the academic champion, the strategic supporter, tutors, students, developers and technologists.

Applications of e-assessment studied included:

- Large scale summative assessment
- University-wide formative e-assessment
- Confidence based testing
- e-Assessment in the Science and Mathematics domain
- e-Assessment being offered to large numbers of distance learners
- Mobile technology input to e-portfolio
- e-Assessment for continuing professional development
- Large scale e-assessment for the general public
  i.e. British Citizenship test
- University entrance test produced by a public examination board

Sites of interest relating to these themes are illustrated in Table 1.

Case Study	e-Assessment Practice	Location & Type of Institution
1. Derby	Large scale summative assessment using the TRIADS assessment engine	HE England
2. The Open University	Numerous bespoke products including OpenMark and OpenMentor	HE England
3. Birkbeck	Assists part time students where English is not their first language. Feedback essential feature	HE England
4. Warwickshire	Assessment at work for the equine industry	FE England
5. West Suffolk	Mobile technology used to collect photographic evidence for e-portfolios on a Chef's course	FE England
6. Dundee	Staff development for quality in the delivery of e-assessment on a university-wide basis	HE Scotland
7. COLA Project	FE staff developed e-assessment questions for a repository used throughout Scotland	FE Scotland
8. Cardiff	Formative e-assessment in oral pathology	HE Wales
9. Coleg Sir Gar	IT for Business course developed and accredited by Edexcel. Formative and summative assessments	FE Wales
10. Ulster	Question*mark*™ Perception™ used to supplement traditional notes and lectures	HE Northern Ireland
11. East Antrim	e-Assessment in process of becoming established throughout the Institution	FE Northern Ireland
12. Southampton, Plymouth, Loughborough	Commercial systems employed university wide throughout these three Universities	HE England
13. UCL & Glamorgan	Confidence based testing employed to encourage reflective practice for learners	HE England
14. Heriott-Watt/Surrey	Specialists in numeric and algebraic assessments also offer partial credits for answers	HE Scotland/England
15. Cambridge Assessment	Development of University of Cambridge entrance test	Public Exam Board
16. British Citizenship Online Examination	Government product produced by commercial company	e-Assessment for general public
17. Chesterfield Royal Hospital	Continuing professional development in medicines administration for nurses	Professional development

Table 1: e-Assessment sites investigated by the project

The e-assessment practices investigated at all of the sites were considered to be wholly or partially distinct from traditional paper and pencil tests with the majority of cases highlighting that they had broken new ground, either from a technical perspective or in their design to solve a learning or learning distribution problem. Four of the applications studied were new to the learning community. The majority of sites reported that they had seen an improvement in student results with their e-assessment applications, whilst over half declared an improvement in retention rates, with the introduction of new e-assessment practices.

One important finding was that the main driver for developing e-assessment was the prior identification of a real pedagogical need. For example at West Suffolk College, students on the chef's course are required to complete an electronic portfolio. For the last two years mobile technologies have been used by students and tutors to record evidence at their place of work such as photographs and video clips for e-portfolios. The e-portfolio used was supplied by Paperfree Systems Ltd (http://www.paperfree.co.uk/).

Students use their mobile phones to send pictures of their culinary creations produced in their working environment to their teachers at the college; these images were also included in their portfolios. Both tutors and students have reported that the teaching has become 'more alive' and this type of assessment has also assisted students previously labelled as under-achievers (because their writing skills were weak) to become more engaged.

The course has been able to demonstrate an increase in retention and achievement since using this form of e-assessment. It has increased student motivation. This model of assessment is attracting attention from the Performing Arts and other vocational courses at the College.

Other factors found to be significant in establishing widespread adoption of e-assessment throughout an institution included active support at senior institutional management level coupled with strong staff development and pedagogical and technical support for tutors from central services.

One of the barriers to expanding e-assessment practice identified by this study is the time and expertise required to develop innovative questions. The COLA project has addressed this issue by forming a consortium of colleges in Scotland whereby learning content is developed and shared (Sclater & MacDonald, 2004). One of the main drivers for adoption of e-assessment in this case study was the availability of funding. The Funding Council had given monies to all of the colleges to buy virtual learning environments and there was a general view that online assessment would encourage staff to use the VLEs more.

This research also uncovered different ways of employing confidence level testing. Glamorgan in their MCQ tests require the students to indicate the level of confidence in the correctness of their answer they have selected. This gives the tutors some way of identifying misconceptions which can then lead to pedagogical changes. Another use has been identified, at UCL, for the classes in medical diagnosis. Here the aim is to assist the students in building their 'aura of confidence' with their professional expertise.

Barriers to cross-institutional adoption of e-assessment included organisational structures that favoured autonomous academic departments, coupled with limited centralised support. Whilst such organisational structures may favour innovative developments, within these pedagogically tight and discipline-focused departments the potential for wider dissemination of the e-assessment methodology across the institution is more restricted.

The role of formative assessment and its effect upon teaching and learning was raised by a number of interviewees. Its advantages were stressed by Birkbeck in their Molecular Cell Biology course. They developed e-assessment (using the TRIAD engine, Mackenzie et al, 2004) to particularly assist students returning to Higher Education after a long leave of absence to gauge their own learning progress. This group had deliberately designed the course materials to encourage deeper learning for the students. The team also showed in their 2001 study that frequent use of computer based assessments were especially beneficial to the many students on programmes where English was not their native language (Baggott & Rayne, 2001). Patterns of use of formative Computer Based Assessments (CBAs) have also been examined and students surveyed (using validated questionnaires) to gain an understanding for the development team of the effect of this type of CBA on study behaviour. These analyses suggest that 'student learning' benefits from the type of e-assessment approach that has been adopted by Birkbeck. Students also appreciated this type of assessment and one student explicitly stated:

"These tests help you reflect upon what you don't know"

and

"…not a boring way to learn"

The British Citizenship Online test which was commissioned by the Home Office is being used by about 60,000 applicants each year. The advantages identified by the commercial group who produced this assessment, was that a tried and tested assessment engine had been chosen to deliver the project. Therefore the project did not start from scratch and could be delivered on time. The driver for adoption of e-assessment for this particular application was that research had shown that learners are more responsive and less nervous if the test is delivered online. e-Assessment is also open to more statistical analysis and its objectivity can be demonstrated easily.

Conclusions

The variety of applications of e-assessment studied and their innovation and general effectiveness indicate the potential of e-assessment to significantly enhance the learning environment and the outcomes for students, in a wide range of disciplines and applications.

The studies illustrate that the principal facilitators for effective implementation of e-assessment include active institutional support from senior management with strong staff development, pedagogical and technical support for tutors from central services. The role of pedagogically sound, imaginative design for e-assessment on the part of tutors is often a significant factor in its success.

Drivers for adoption of e-assessment included perceived increases in student retention, enhanced quality of feedback, flexibility for distance learning, strategies to cope with large student/candidate numbers, objectivity in marking and more effective use of virtual learning environments.

The principal barrier to development of institution-wide e-assessment remains one of academic staff time and training. Dissemination of pockets of innovative e-assessment practice across an institution can be hampered by organisational structures that favour autonomous academic departments, and limited centralised support.

References

Baggot, G. and Rayne, R.C. (2001) Learning support for mature, part-time, evening students: providing feedback by frequent, computer-based assessments. In Proceedings of the 5th International CAA Conference.

M.Danson and C. Eabry (eds) Loughborough University pp.9-20. Available online at http://www.lboro.ac.uk/service/ltd/flicaa/conf2001/pdfs/contents.pdf

Gomm, R., Hammersley, M. and Foster, P. (2000) (eds) Case Study Method, Key Issues, Key Texts. Sage Publications. ISBN 0 7619 6414 2

Mackenzie, D.M., O'Hare, D., Paul, C., Boyle, A., Edwards, D., Williams, D. and Wilkins, H. (2004) Assessment for learning: the TRIADS Assessment of Learning Outcomes Project and the Development of a pedagogically-friendly computer-based assessment system. In O'Hare, D. and Mackenzie D. (eds) Advances in Computer Aided Assessment, SEDA Paper 116 pp.11-24. Staff & Educational Development Association Ltd, Birmingham. ISBN 1 902435 24 9.

Mackenzie, D.M. (2005) Online Assessment: quality production and delivery for higher education in Enhancing Practice, Keynote Address, Assessment Workshop Series No. 5 in Reflections on Assessment, Volume II pp.22-29, Quality Assurance Agency for Higher Education, Gloucester. ISBN 1 84482 266 4.

Sclater, N. & MacDonald, M. (2004) Developing a National Item Bank. Proceedings of the 8th International CAA Conference, Loughborough, July.

The 14-19 Education & Skills White Paper presented to Parliament by The Secretary of State for Education and Skills by Command of Her Majesty in February 2005. Last updated 10th October 2005. Author: Ruth Kelly

Whitelock, D. (In Press) Electronic assessment: marking, monitoring and mediating learning. In McAndrew, P. and Jones. A. (eds) Interactions, Objects and Outcomes in learning. Special Issue of International Journal of Learning Technology.

Yin, R.K. (2003) Applications of Case Study Research, Second Edition, Applied Social Research Methods Series, Vol. 34, Sage Publications. ISBN 0 7619 2551 1 http://www.paperfree.co.uk/ Paperfree Systems Ltd – e-Portfolio system.

R2Q2: RENDERING AND REPONSES PROCESSING FOR QTIV2 QUESTION TYPES

Gary Wills, Hugh Davis, Swapna Chennupati, Lester Gilbert, Yvonne Howard, Ehtesham-Rasheed Jam, Steve Jeyes, David Millard, Robert Sherratt and Gavin Willingham

R2Q2: Rendering and Reponses Processing for QTIv2 Question Types

Gary Wills[‡], Hugh Davis[‡], Swapna Chennupati[‡], Lester Gilbert[‡], Yvonne Howard[‡], Ehtesham-Rasheed Jam[‡], Steve Jeyes[†], David Millard[‡], Robert Sherratt[†] and Gavin Willingham[‡]

[‡]Learning Technologies Group, University of Southampton, UK.
[†]e-Services Integration, University of Hull, UK

Abstract

IMS QTI is a popular and important standard for e-learning assessment. The second version of the standard (QTIv2) works alongside other IMS standards, but take-up has been slow, with problematic implementations and no definitive reference software. The R2Q2 project aims to produce a set of loosely coupled web services that will provide definitive reference software for QTIv2. In this paper we describe how we have learnt from previous development efforts in order to produce a first architecture and initial implementation. We also describe the results of our interviews with the wider QTI community to identify what is believed to be important for our planned final reference implementation.

Introduction

E-learning assessment covers a broad range of activities involving the use of machines to support assessment, either directly (such as web-based assessment tools, or tutor systems) or indirectly by supporting the processes of assessment (such as quality assurance processes for examinations). It is an important and popular area within the e-learning community [6, 1, 2] Within this broad view of e-learning assessment, the domain appears established but not mature, as traditionally there has been little agreement on standards or interoperability at the software level. Despite significant efforts by the community, many of the most popular software systems are monolithic and tightly coupled, and standards are still evolving.

One of the more popular standards that has emerged is Question and Test Interoperability (QTI) developed by the IMS Consortium[1]. The QTI specification describes a data model for representing questions and tests and the reporting of results, thereby allowing the exchange of data (item, test, and results) between tools (such as authoring tools, item banks, test

[1] **IMS QTI homepage: http://www.imsglobal.org/question/**

constructional tools, learning environments, and assessment delivery systems) [10]. Wide take-up of QTI would facilitate not only the sharing of questions and tests across institutions, but would also enable investment in the development of common tools. QTI is now in its second version (QTIv2), designed for compatibility with other IMS specifications, but despite community enthusiasm there have been only a few real examples of QTIv2 being used, and no definitive reference implementation [8,9].

In the last few years there has been a trend away from tightly coupled monolithic systems towards Service-Oriented Architectures (SOA). SOAs are an attempt to modularise large complex systems in such a way that they are composed of independent software components that offer services to one another through well-defined interfaces.

One way to promote QTIv2 is through a reference implementation of the standard written within the service-oriented paradigm. In the UK, the Joint Information Systems Committee (JISC) is financed by all the Further and Higher Education funding councils within the country, and is responsible for providing advice and guidance on the use of Information and Communications Technology (ICT) for learning and teaching. Part of their strategy is the development of a SOA framework for e-learning [5,7], and of reference models that describe how different areas of e-learning can be supported by the framework.

For the assessment domain, the reference model is FREMA (Framework Reference Model for Assessment)[2]. The FREMA project has defined a number of high level service profiles that describe how services can work together within the assessment domain to fulfil particular use cases [4]. Several of these use cases require questions to be rendered, answers taken, and feedback to be generated. The corresponding services provide an ideal opportunity to create a reference implementation of the core functionality of QTIv2 that fits within the broader FREMA context.

This paper will report on the progress of the R2Q2 project. R2Q2 is a JISC funded project that aims to bring the SOA approach and QTI standard together to develop a set of Web Services that will render and respond to questions written in the QTI standard.

Service Oriented Architectures

A service approach is ideally suited to more loosely coupled systems, where individual parts may be developed by different people or organizations. Wilson *et al.* [7] discuss in detail the advantages of using a SOA:

[2] **FREMA homepage: http://www.frema.ecs.soton.ac.uk/**

- **Modularity**: As services are dynamically coupled, it is relatively easy to integrate new services into the framework, or exchange new implementations for old.

- **Interoperability**: Due to standardization of the communication and description of the services, third party services can easily be incorporated as required.

- **Extensibility**: Due to the relative ease with which services can be incorporated into a system, there is less danger of technology 'lock-in'.

Due to the nature of the loose coupling in a SOA, applications can be developed and deployed incrementally. In addition, new features can be easily added after the system is deployed. This modularity and extensibility make SOA especially suitable as a platform for an assessment system with evolving requirements and standards. Services are also appealing in terms of their ability to be reused, as they have well-defined public interfaces. In R2Q2 we will be developing web services that are built on widely used standards such as SOAP and WSDL. It is our hope that this will make it easy for other members of the community to use the services, and further develop them.

Question and Test Interoperability

The IMS QTI Specification is a standard for representing questions and tests with a binding to the eXtended Markup Langage (XML, developed by the W3C) to allow interchange. Figure 1 shows a short example of a question expressed in this format, taken from the IMS QTI examples. This example is a simple multiple choice question, illustrating the core elements: *ItemBody* declares the content of the question itself, *ResponseDeclaration* declares a variable to store the student's answer, and *OutcomeVariables* declares other resulting variables, in this case a score variable to hold the value of the result.

In R2Q2 we focus on rendering and responding to the 16 different types of interactions described in version 2 of the QTI specification (QTIv2). These are:

1) Choice
2) Order
3) Associate
4) Match
5) Inline Choice
6) Text Entry
7) Extended Text
8) Hot Text
9) Hotspot
10) Select point
11) Graphic
12) Graphic Order
13) Graphic Associate
14) Graphic Gap Match
15) Position object
16) Slider

```
<?xml version="1.0" encoding="UTF-8"?>
<assessmentItem xmlns="http://www.imsglobal.org/xsd/imsqti_v2p0"
    identifier="choice" title="Unattended Luggage"
    adaptive="false" timeDependent="false">
    <responseDeclaration identifier="RESPONSE" cardinality="single"
                         baseType="identifier">
        <correctResponse>
            <value>ChoiceA</value>
        </correctResponse>
    </responseDeclaration>
    <outcomeDeclaration identifier="SCORE" cardinality="single"
                        baseType="integer">
        <defaultValue>
            <value>0</value>
        </defaultValue>
    </outcomeDeclaration>
    <itemBody>
        <p>Examine the following sign:</p>
        <p>
           <img src="images/sign.png" alt="NEVER LEAVE LUGGAGE UNATTENDED"/>
        </p>
        <choiceInteraction responseIdentifier="RESPONSE"
                           shuffle="false" maxChoices="1">
            <prompt>What does it say?</prompt>
            <simpleChoice identifier="ChoiceA">You must stay with your
                 luggage at all times.</simpleChoice>
            <simpleChoice identifier="ChoiceB">Do not let someone else look
                 after your luggage.</simpleChoice>
            <simpleChoice identifier="ChoiceC">Remember your luggage when
                 you leave.</simpleChoice>
        </choiceInteraction>
    </itemBody>
    <responseProcessing template =
    "http://www.imsglobal.org/question/qti_v2p0/rptemplates/match_correct"/>
</assessmentItem>
```

Figure 1: Example QTIv2 question (abridged for simplicity)

The list of different question types can be combined with templated question or adaptive response, providing an author with numerous alternative methods for writing questions appropriate to the needs of the students. Templated questions include variables in their item bodies that are instantiated when a question is rendered (for example, inserting different values into the text of maths problems). Adaptive questions have a branching structure, and the parts that a student sees depends on their answer to each part of the branch. In total these allow for sixty four different possible combinations.

Previous Work

One of the earliest successful projects in the area of rendering and response using the QTI standard was the Assessment Provision through Interoperable Segments project (APIS) [8]. This was later reused in the ASSIS project as a core service, called *QTIRun* [9]. The APIS project aimed to implement a modular item rendering engine in line with QTIv2. Whilst the APIS and ASSIS

projects have provided a launch pad from which many other projects have benefited, there are a number of short-comings in their final implementations.

- *QTIRun* is implemented as a single Web Service, and in order to preserve the statelessness of the render/response functions, the service calls pass excessively large amounts of XML data around the system.

- Despite this the interactions between services remained tightly coupled, compromising extensibility, such that if a different render engine was required (or a different response engine) the code would have to be re-written.

- A lack of documentation has resulted in confusion over the type of QTIv2 questions served by *QTIRun*. In fact the *QTIRun* service only deals with a limited subset of QTIv2 question types.

The aim of the R2Q2 project is to learn from the experiences of APIS and ASSIS and produce a genuinely loosely coupled SOA for flexibility and extensibility. The project uses an agile software engineering methodology in which every stage is carefully documented, the main points of which are published weekly on the project website in the form of a blog.

R2Q2 Design

The first stage of the design was to examine what had been built before in the APIS project and identify the lessons described above. Also in the initial stages of the project we interviewed people outside of the project team who are actively developing Web services for assessment and/or developing the QTI specification. An overview of the services the R2Q2 system will provide is shown in Figure 2.

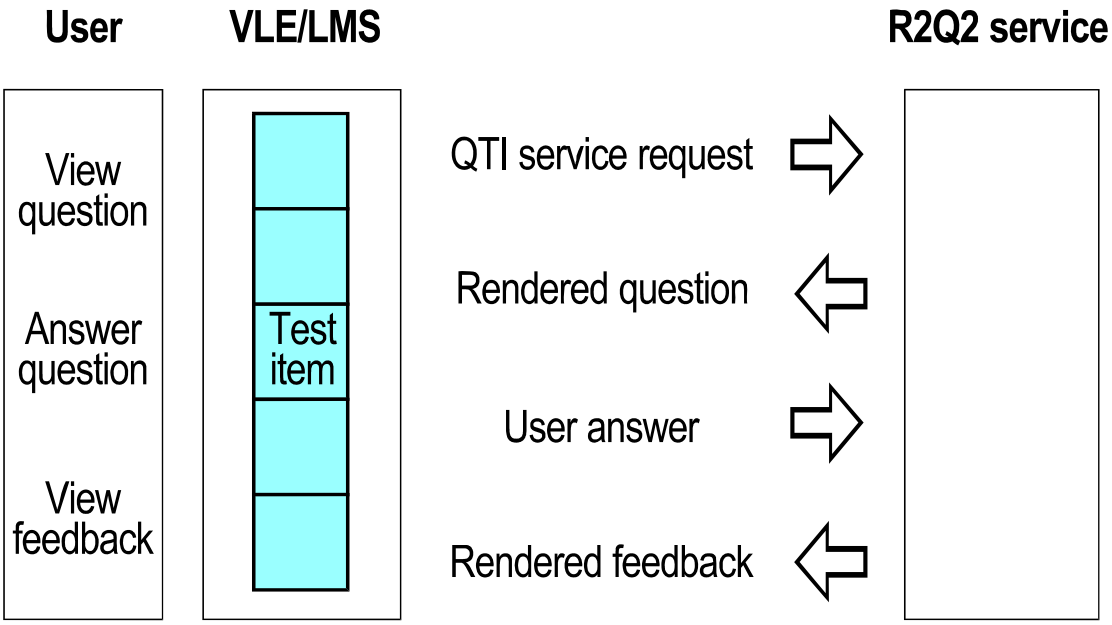


Figure 2 R2Q2 Overview

In the R2Q2 project we aim to provide a service that is more reliable than *QTIRun*, with definitive render and response processing engines for QTI version 2 question types. This is achieved by taking the single *QTIRun* Web Service and refactoring it such that the main functions are divided between several co-operating Web Services. In our first development iteration we have focused on what we believe are the core functions (see Figure 3), allowing us to extend the service once we have validated the system.

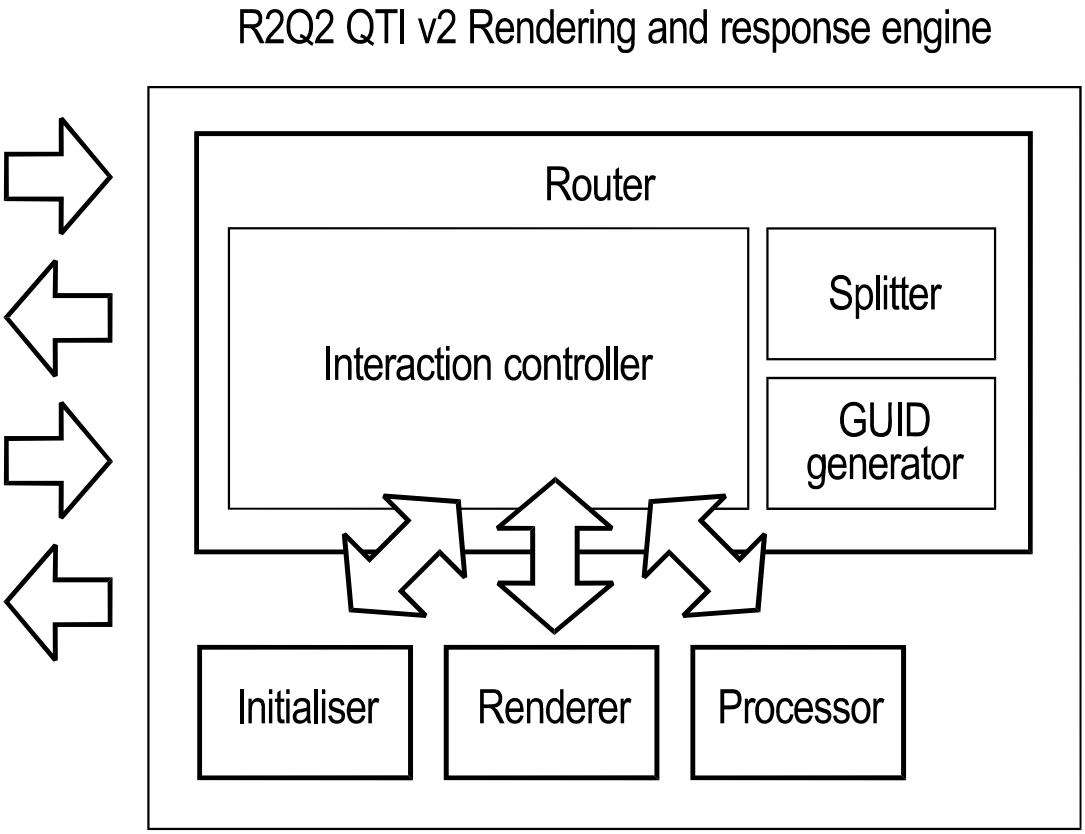


Figure 3 The R2Q2 Architecture

The R2Q2 engine is a loosely coupled architecture comprising of three interoperable services. All the interactions with and within the R2Q2 engine are managed by an internal component called the Router.

The Router is responsible for parsing and passing the various components of the item (QTIv2) to the responsible web services. It also manages the interactions of external software with the system, and it is therefore the only component that handles state. This enables the other services to be much simpler than QTIRun, and they can maintain a loosely coupled interface but without the need to exchange large amounts of XML.

The Processor service processes the user responses and generates feedback. The Processor compares the user's answer with a set of rules and generates response variables based on those rules. The Renderer service then renders the item (and any feedback) to the user given these response variables.

Future Development

Figure 3 shows the core services where R2Q2 is used as a stand alone service. However, R2Q2 is also designed to be dropped into applications such as a test engine or authoring tool. The second iteration of the design will therefore develop the services that will allow the R2Q2 engine to be integrated into other community projects. From the interviews we have conducted there are several areas that the interviewees felt needed attention:

- Authors would like to be able to batch-process questions and answers.
- While the specification only gives examples in XHTML, it would be good to have a rendering process for questions using Flash.
- Some management of service loading and subsequent performance is required (as many users may attempt to take a given test at the same time).
- The use of the Remote Question Protocol (RQP) needs investigation as it may allow R2Q2 to be easily integrated into a VLE such as Moodle.
- Good documentation is essential if this tool is to be used by others.
- A single install process is important for community take-up of any tool.

Conclusions

At a recent conference the UK assessment community confirmed that kick-starting the use of the IMS Question and Test Interoperability version 2 specifications was a high priority. Whilst earlier versions of the specification provided most of the functions needed by practitioners, to ensure future interoperability it was considered essential that tools migrate to this new standard. However there was little incentive to move towards the new specification as existing public implementations are incomplete. The conference concluded that there needed to be a robust set of tools and services that conformed to the QTIv2 specification to facilitate this migration.

A central function that many systems require is that of rendering questions and responding to users answers. The R2Q2 project aims to produce a core set of web services to provide this functionality.

To ensure wide-spread take up of the specification, however, R2Q2 will need to be integrated into authoring tools, test engines, VLEs and LMSs, amongst other applications, to achieve the aim of migrating the community to this new standard.

References

[1] Bull J., and McKenna C. Blueprint for Computer Assisted Assessment. Routledge Falmer, 2004.

[2] Conole, G. and Warburton, B. "A review of computer-assisted assessment". ALT-J Research in Learning Technology, vol. 13, pp. 17-31, 2005.

[3] Cooper, A. and Reimann, R. About Face 2.0: The Essentials of Interaction Design. John Wiley & Sons, 2003.

[4] Davies, W. M., Howard, Y., Millard, D. E., Davis, H. C. and Sclater, N. Aggregating Assessment Tools in a Service Oriented Architecture. In Proceedings of 9th International CAA Conference, Loughborough. 2005

[5] Olivier B., Roberts T., and Blinco K. "The e-Framework for Education and Research: An Overview". DEST (Australia), JISC-CETIS (UK), www.e-framework.org, accessed July 2005.

[6] Sclater N., Howie K. User requirements of the ‘‘ultimate’’ online assessment engine, Computers & Education, 40, 285–306 2003.

[7] Wilson S., Blinco K., and Rehak D. Service-Oriented Frameworks: Modelling the infrastructure for the next generation of e-Learning Systems. JISC, Bristol, UK 2004.

[8] APIS [Assessment Provision through Interoperable Segments] - University of Strathclyde–(eLearning Framework and Tools Strand) http://www.jisc.ac.uk/index.cfm?name=apis, accessed 30 April 2006.

[9] Assessment and Simple Sequencing Integration Services (ASSIS) – Final Report – 1.0. http://www.hull.ac.uk/esig/downloads/Final-Report-Assis.pdf, accessed 29 April 2006.

[10] IMS Global Learning Consortium, Inc. IMS Question and Test Interoperability Version 2.1 Public Draft Specification. http://www.imsglobal.org/question/index.html, accessed 9 January 2006.

IMPLEMENTING LARGE SCALE ASSESSMENT PROGRAMMES

Liam Wynne and Suzana Lopes

Implementing large scale assessment programmes

Liam Wynne and Suzana Lopes
Pearson Vue
Liam.Wynne@pearson.com
www.pearson.com

More and more testing organisations are moving their traditional paper and pencil assessment programmes to computer-based testing. There is little doubt that computer-based testing offers many advantages, but careful consideration needs to be given to the transition from paper and pencil to computer.

Converting to computer-based testing can streamline back end business functions and cut costs by reducing administration and improving efficiencies. It is important at the business requirements stage of the transition to look at what impact this new way of delivering tests will have on the business and how it may be able to improve operational efficiency, existing processes and procedures.

Another consideration is what the final test will look like? Will it be just a straight conversion to computer of the paper based test or will it be a more complex computer adaptive test? When first converting to computer-based testing many organisations start with a straight conversion of the paper based test and then progress to more complex forms of testing.

Testing organisations also need to determine whether to deliver their tests online or via a server. This is largely dependent on the environment in which the test will be delivered and the level of security required. Low stakes or practice tests can be delivered directly over the internet, whereas most high stakes testing organisations would prefer the security of delivering their tests via a server.

There are also administrative procedures to take into consideration. Where will the test be delivered? Does the testing organisation's current channel have the necessary IT infrastructure to support computer-based delivery? Is the integrity of the test at risk using the current channel due to lack of security? Are there other potential markets that could be explored by expanding the existing channel?

If there is a charge for the test can the testing and registration system take into account different pricing structures? Can it accept credit card payment and manage invoices? Can the testing organisation take accounting control for customers and invoice or them on a monthly basis? Perhaps it is more cost effective to outsource financial services to a testing partner?

With computer-based testing it is much easier to get useful management information compared to paper and pencil tests. This can help improve both the test and marketing as well as provide feedback to candidates on how to improve their learning. When converting to computer-based testing, organisations need to think about what data is going to be useful for growth in the future, who else needs access to this data and what reports will be required.

Of critical importance for large scale or rapid growth assessment programmes is the testing infrastructure. This is the glue that sticks the test development, delivery, financial services, registration and scheduling and information management together. The infrastructure needs to be reliable and scalable, so that as the volume of tests delivered increases the system is able to support this increase.

This presentation will consider the experience of two of the largest testing organisations in the UK. It will review their business objectives for converting to computer-based testing, critical success factors and implementation timescales.