

<http://www.caaconference.com>

# 11<sup>th</sup> CAA

---

## INTERNATIONAL COMPUTER ASSISTED ASSESSMENT CONFERENCE

---

Proceedings of the Conference  
on 10th & 11th July 2007  
at Loughborough University

Edited by Farzana Khandia

*Research into E-Assessment*

Published by: Professional Development  
Loughborough University  
Loughborough  
Leicestershire  
LE11 3TU  
UK

Tel: +44(0)1509 222893

Fax: +44(0)1509 223992

<http://www.lboro.ac.uk/service/pd/>

ISBN: 0-9539572-6-8

© Loughborough University

## Commercial Sponsors

We are pleased to thank the following companies and organisations for their generous sponsorship of the 2007 CAA Conference.



[www.questionmark.com](http://www.questionmark.com)



[www.pearsonvue.com](http://www.pearsonvue.com)



[www.i4learn.co.uk](http://www.i4learn.co.uk)



[www.techdis.ac.uk](http://www.techdis.ac.uk)





# Contents

---

## Foreword

Conference Director, Loughborough University

## Advisory Committee and Reviewers

## Papers

<b>Al-Amri S</b> Computer-based vs. paper-based testing: are they the same?	3
<b>Armellini A Jones S Salmon G</b> Developing Assessment for Learning through e-Tivities	13
<b>Ball S</b> Accessibility in E-Assessment	21
<b>Barker T Lee S</b> The verification of identity in online assessment: a comparison of methods	37
<b>Baruah N Greenhow</b> Exploring the Potential, Limitations and use of Objective Questions in Advanced Calculus	47
<b>Bedford S Price G</b> A Study into the use of Computer Aided Assessment to Enhance Formative Assessment during the early stages of Undergraduate Chemistry Courses	57
<b>Bertolo E Lambert G</b> Implementing CAA in chemistry: a case study	73
<b>Boyle A</b> The Formative use of e-Assessment: some early implementations, and suggestions for how we might move on	87
<b>Boyle A</b> Principles for the regulation of e-assessment: An update on developments	111
<b>Chew E Jones N</b> The Marriage of Freire and Bloom: An Assessment Prototype for Pedagogy of the Oppressed and Higher Order Thinking	117
<b>Cowan P</b> Using CAT for 11-Plus Testing in Northern Ireland: What are the issues?	129
<b>Craig N</b> What's New in e-Assessment? From Computer-Marking to Innovative Item Types	139
<b>Davies P</b> Review of the Computerized Peer-Assessment of Essays. Will it have an Effect on Student Marking Consistency?	143

<b>Dermo J</b> Benefits and Obstacles: Factors Affecting the Uptake of CAA in Undergraduate Courses	155
<b>Ekins J</b> The use of Interactive On-line Formative quizzes in Mathematics	163
<b>Elliott B</b> Modernising assessment: the use of Web 2.0 for formative and summative assessment	179
<b>Fluck A</b> Can Tertiary e-Assessment Change Secondary School Cultures?	191
<b>Gomersall B</b> Assessment and Learning: Is Assessment an Afterthought or is it at the Heart of the Learning Process?	207
<b>Guest E Brown S</b> A new Method for Parsing Student Text to Support Computer-Assisted Assessment of Free Text Answers	223
<b>Hackett E Seddon P</b> From Online Entries to Online Results	239
<b>Harrison G Gray J</b> An Improved Computer-Assisted Test for Accessible Computer-Assisted Assessment	253
<b>Hepplestone S Mather R</b> Meeting Rising Student Expectations of Online Submission and Online Feedback	269
<b>Heron C</b> A Hardware Solution for Access to CAA for Students with Reduced Manual Dexterity due to Acute Neurodisability – A Case Study	279
<b>Jones S</b> Diagnosing and Developing the IT Skills of New Entrants to Higher Education	289
<b>MacKenzie D Stanwell M</b> QuickTrl A Visual System for the Rapid Creation of e-Assessments and e-Learning Materials	301
<b>McAlpine M Glauert J Jennings V Thomas N Rabin A</b> Incorporating Avatar Signing Into Assessment Items	305
<b>McAlpine M</b> The use of Wikis for Assessing Collaborative Learner Achievement	311
<b>McLaughlin P Kerr W Howie K</b> Fuller, Richer Feedback, More Easily Delivered, using Tablet PCs	329
<b>Meyer J Ziman M Fyfe S Fyfe G Plastow K Sanders K Hill J</b> Implications of Patterns for use of Freely Available Online Formative Tests for Summative Tasks	343

<b>Ruedel C Whitelock D Mackenzie D</b> Key Factors for Effective Organisation of e-Assessment	357
<b>Scheuermann F Guimarães Pereira Â</b> Quality aspects of Open Source Testing Tools	371
<b>Schmid F Mitchell T Whitehouse J Broomhead P</b> EXAMONLINE: e-Enabling “Traditional” HE/FE Examinations	381
<b>Shepherd E</b> Blended Delivery Meets e-Assessment	399
<b>Tselonis C Sargeant J</b> Domain-Specific Formative Feedback through Domain-Independent Diagram Matching	403
<b>Whitelock D Watt S</b> Open Mentor: Supporting Tutors with their Feedback to Students	421
<b>Williams J Bedi K</b> Using Digital Storytelling as an Assessment Instrument: Preliminary Findings at an Online University	433
<b>Wills G Bailey C Davis H Gilbert L Howard Y Jeyes S Millard D Price J Sclater N Sherratt R Tulloch I Young R</b> An e-Learning Framework for Assessment (FREMA)	451
<b>Wills G Davis H Gilbert L Hare J Howard Y Jeyes S Millard D Sherratt R</b> Delivery of QTIv2 Question Types	473
<b>Yu X Lowe J</b> Computer Assisted Testing of Spoken English: A Study to Evaluate the SFLEP College English Oral Test System in China	489



## Foreword

Welcome to the eleventh International CAA Conference at Loughborough University.

The CAA Conference provides an opportunity to bring together professionals from across the world who are interested in CAA related research and findings. Within these Proceedings you will find contributions from Awarding Bodies, Higher Education, Research Committees, etc.

This year it contains thirty nine papers which have passed our double blind refereed process. Our thanks go to the authors of all the papers that were submitted. The range and depth of their research is ample justification for a conference in the field of CAA. We also acknowledge the time and effort contributed by the Advisory Panel; their help has been invaluable in shaping the conference and maintaining its standards.

In response to the positive feedback regarding the themed programme of last year, we have loosely themed the programme again this year. A number of topics emerge from the papers, a mixture of both a technical and pedagogic nature. One is a growing interest in Web 2.0; even more relevant is the topic of assessing using Web 2.0 technologies, and there are a couple of papers addressing this.

The Joanna Bull Prize for best paper will be announced at the event and publicised on the conference website ([www.caaconference.co.uk](http://www.caaconference.co.uk))

Once again we welcome our commercial sponsors. This year Questionmark Computing are joined by i4L, JISC Techdis and Pearson VUE all of whom will be exhibiting in the display area as well as presenting - do please pay them a visit.

Lastly, a conference isn't a conference without delegates and this year we see them come from far and wide. We do hope you all find the conference a valuable and worthwhile experience. As in previous years, we encourage our repeat attendees to engage with those new to the CAA Conference and benefit from the two days both professionally and socially.

Enjoy the conference.

Farzana Khandia  
July 2007



## 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

**Clive Church**, Edexcel

**Graeme Clark**, Adam Smith College

**Grainne Conole**, University of Southampton

**Phil Davies**, University of Glamorgan

**Bobby Elliott**, Scottish Qualifications Authority

**Graham Farrell**, Swinburne University of Technology

**Don MacKenzie**, University of Derby

**Mhairi McAlpine**, Scottish Qualifications Authority

**Nora Mogey**, University of Edinburgh

**John Sargeant**, University of Manchester

**Derek Stephens**, Loughborough University

**Bill Warburton**, University of Southampton

**Denise Whitelock**, Open University

**Jeremy Williams**, U21Global

**Rowin Young**, University of Strathclyde





**COMPUTER-BASED  
VS.  
PAPER-BASED TESTING:  
ARE THEY THE SAME?**

**Saad Al-Amri**



# **Computer-based vs. Paper-based Testing: Are They the Same?**

Saad Al-Amri  
University of Essex, UK ( [salamr@essex.ac.uk](mailto:salamr@essex.ac.uk) )

## **Abstract**

Comparability studies of paper-based tests and computer-based tests focused mainly on the equivalence of both forms and the contributing factors affecting this concept. There have been several studies in different fields such as education, psychology, ergonomics and L1 reading research (Sawaki, 2001). However, there has been no empirical study so far that has investigated the effect of prior computer familiarity on students' performance taking L2 tests (Chapelle & Douglas, 2006). Chapelle & Douglas (ibid) also mentioned the significance and lack of differential validity studies and how motivating it is to find out more about performance on computer-based testing. Sawaki (2001) argues that this type of empirical work should employ different methodologies such as eye movement, verbal protocols, post hoc interviews, and questionnaires to reveal useful results. Thus, this ongoing study examines the comparability of paper based tests and the computer based testing in L2 reading context, and the impact of test takers' characteristics, i.e., computer familiarity, computer preference, gender and test taking strategies on students' performance on computer based tests, and sequentially on comparability with paper based tests. 167 Saudi medical students took three reading achievement tests on both paper and computer. The Test of English as a Foreign Language (TOEFL) was devised to measure the students' proficiency and anchor the study tests. The study questionnaires focused on demographic information, participants' computer familiarity and preference of testing mode. The interview examined any change of preference after exposure to CBT. A triangulation of think aloud reports and post hoc interviews were employed to gain insight into strategies used on both testing modes, and to confirm comparability of both modes for validity purposes. The results are likely to reveal some information about the equivalence of both testing modes based on a scientific systematic perspective and have implications for the implementation of computer based reading tests in the context of the medical faculty EAP programme.

## **Introduction and Literature Review**

Technological advancement has moved very rapidly since the last century. Computers became the most useful facilitator in achieving the majority of our goals. Technology has been implemented in the field of language assessment by using computers to deliver different types of assessment. However, little empirical work has been done in order to examine the influence of technology

on the essence of the assessment, which is the concept of validity. Moreover, little research has been conducted to investigate the interaction between the assessment mode and the test takers. There have been some studies that have focused on the comparability of the paper based testing and the computer based testing in some areas such as psychology, mathematics, and ergonomics (Sawaki, 2001). Yet, only a few studies have looked at this issue in the field of language assessment, such as those done by Taylor, et al (1999); Kirsch, et al (1998); Taylor, et al (1998); Eignor, et al (1998); Russell (1999); Russull and Haney (1997); and Choi, et al (2003). Some studies have revealed that there is a significant difference between the two testing modes (Pomplun, 2002; Choi, et al., 2003) while others have concluded the opposite (Boo, 1997; Whitworth, 2001; Bugbee, 1996).

However, previous research has mainly focused on the product, i.e., test scores achieved, and neglected the process that resulted in these scores. Paek (2005) asserted the significance of computer familiarity, test taking strategies and academic subjects in measuring the equivalence of the two testing modes. Therefore, this study aims to measure the equivalence of the computer based and paper based testing, and consequently, the influence on the essence of the assessment, which is construct validity as defined by Messick (1989). This study also examines how test taker characteristics such as computer familiarity, preference, gender and test taking strategies interact with the testing mode, and to what extent this interaction affects the test scores and as a result the overall construct validity. The methodology used in this study differs from the previous research as the framework employed here is both quantitative and qualitative in nature. This framework triangulates the data sources to increase the validity and reliability of the results and conclusions of this study.

Evaluating the comparability and equivalence of both paper based and computer aided assessment is very crucial before introducing computer aided assessment into any context. It is always vital to ensure that both test qualities and test takers are not disadvantaged by shifting from one mode to another. It is also necessary to survey the readiness of the target context when deciding to implement the computer aided assessment. Conclusions and recommendations of this study will be of interest to the medical colleges in Saudi Arabia as they are the target audience in the target context.

## **Research Questions**

This study attempts to answer the following questions:

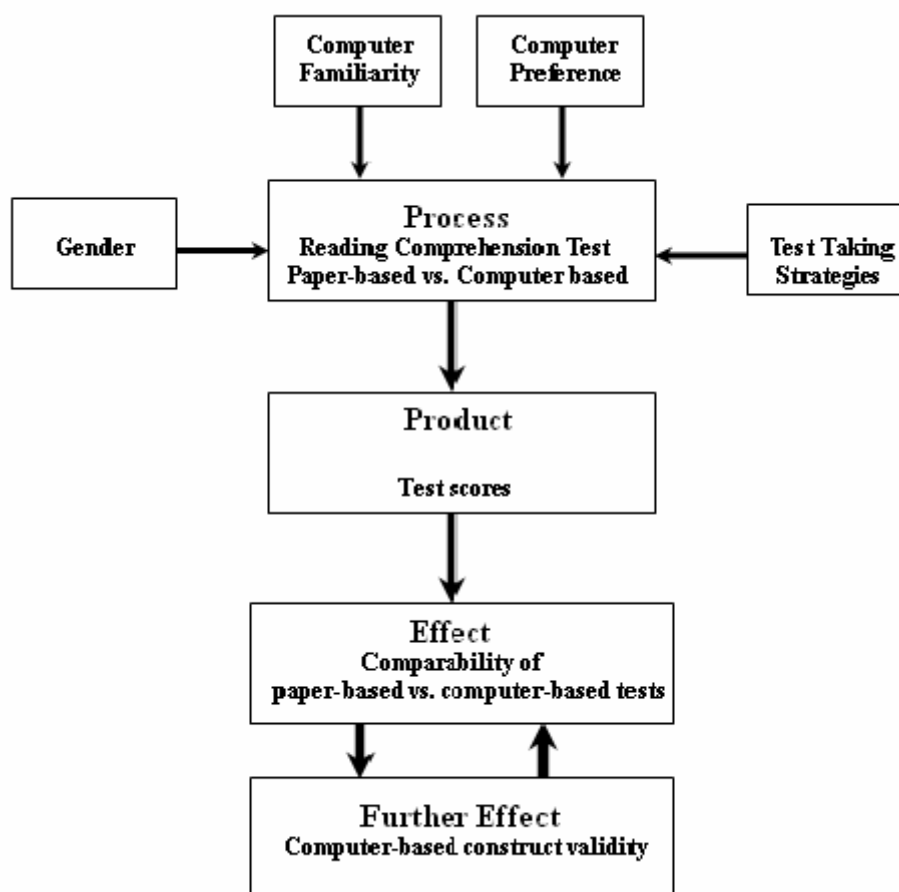
- Would construct validity be influenced by the test administration mode?
- Who will perform better on computer based tests: male or female and why?
- To what extent will students' performance change after exposure to a series of computer based tests?

- To what extent does computer familiarity affect students' performance on computer based tests?
- Do students use the same test taking strategies when taking the same test in two different testing modes?

## Methodology

This study exploited the triangulation perspective in varying the sources of data collection. Thus, a range of quantitative and qualitative instruments were employed to gather the data required for this research. The following diagram explains the framework used in this study:

### Study Framework



There were four instruments used to collect data for this project. They are as follows: the TOEFL test was used to measure the students' proficiency and to anchor the study tests. The study used two questionnaires that were built and designed based on questionnaires from the literature. The construction of the two questionnaires relied on adaptation of some existing questionnaires as well as implementing some new items to suit the study aims. Questionnaire one was designed to collect demographic data, measure computer familiarity

of participants, and computer preference of testing mode. The second questionnaire was used to collect data about preference after exposure to computer aided assessment. There were four computer assessment practical questions that were constructed to measure participants' computer familiarity in accordance with their responses on the first questionnaire. This aims at increasing the accuracy of measuring this construct. The study made use of three institutional achievement tests as a tool to compare the scores on both testing modes. These tests were converted into computer versions using the QuestionMark system. This tool measured score equivalence on both administration modes. Think aloud protocols were used to gain insight at the test taking strategies used by the participants when doing the same test on two different administration modes. Interviews were employed to collect more data about the participants' preference of CBT and the influence of shifting the testing mode on test taking strategies.

### **Data analysis**

The data has been very recently collected and the coding and analyses have not been entirely finished. However, a very basic preliminary analysis has been conducted for the purpose of this paper. The full paper and analysis of both quantitative and qualitative data will be ready for publication at the 12th CAA conference.

First, the tests scores analysis showed that there is a difference in students' performance on paper based tests and computer based tests. This difference becomes quite obvious when looking at the mean scores of paper based and computer based tests. Table 1 shows a summary of all tests' mean scores and standard deviations.

**Table 1. Descriptive Statistics**

Tests	N	Mean	Std. Deviation
Score of paper test 1	167	77.45	12.879
Score of paper test 2	167	66.47	17.856
Score of paper test 3	167	67.90	18.362
Score of computer based test 1	167	74.65	15.389
Score of computer based test 2	167	61.89	19.022
Score of computer based test 3	167	64.07	15.259

Second, by running T-test, there is a high significant correlation between the means of the paper based and computer based tests.

### Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Mean of Paper-based Tests	70.61	167	11.916	.922
	Mean of Computer-based Tests	66.87	167	13.215	1.023

### Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Mean of Paper based Tests & Mean of Computer based Tests	167	.738	.000

### Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Individual Mean of Paper based Test - Individual Mean of Computer-based Tests	3.734	9.180	.710	2.332	5.137	5.257	166	.000

This difference cannot be interpreted immediately as the data concerning the independent variables, i.e., computer familiarity, gender, computer preference and test taking strategies, has not been coded and analysed yet.

Nonetheless, a simple calculation of those who have changed their preference before and after exposure to computer based tests can be noticed from these two questions. Question one was included in questionnaire one before exposure to computer based tests while question two was inserted in questionnaire two after the examinees had finished all the study tests; the following tables indicate the change in the participants' preferences:

**( A question asked BEFORE exposure to CBTs )**  
**Would you prefer taking tests?**

		Frequency	Percent
Valid	On paper	68	40.7
	No difference	47	28.1
	On computer	52	31.1
	Total	167	100.0

**( A question asked AFTER exposure to CBTs )**  
**Which test do you prefer taking again?**

		Frequency	Percent
Valid	On Paper	58	34.7
	No Difference	22	13.2
	On Computer	87	52.1
	Total	167	100.0

By comparing the frequencies of those who preferred paper based tests before and after taking computer based tests, we can see that 6% of those participants have changed their preference. Moreover, the percentage of those who do not mind taking the test in both modes dropped dramatically from 28.1% to 13.2%. On the other hand, this shift in preference resulted in an increase of 21% in the participants who favored the experience of computer based tests.

Further qualitative data was collected by interviewing 23 students sampled out of those participants who have changed their preference. These data have not been processed yet; however, we are certain that they will reveal more interesting findings.

## **Conclusion**

This study aims to measure the comparability of both paper based and computer based L2 reading achievement tests. It also investigates the factors affecting the equivalence of both tests. Because the data coding and analyses are still in their early stages, no clear picture can yet be made about the final findings and results of this study. Also, most of the variables of this study cannot be discussed in this paper for the same reason. In addition, final solid conclusions cannot yet be drawn out of the available results due to the limited time available for analysis. This study is still ongoing and it is hoped that by the 12<sup>th</sup> CAA conference, the complete results and findings, as well as the entire project, will be ready for publication.



## References

- Boo, J (1997). *Computerized versus paper-and-pencil assessment of educational development: Score comparability and examinee preferences*. Unpublished dissertation, University of Iowa.
- Bugbee, A. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education*, vol. 28 (3) pp. 282-300.
- Chapelle, C. & Douglas, D., (2006). *Assessing language through computer technology*. Cambridge. UK: CUP.
- Choi, I., Kim, K. and Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing*, vol. 20 (3) pp. 295-320.
- Eignor, D., Taylor, C., Kirsch, I. and Jamieson, J. (1998). Development of a scale for assessing the level of computer familiarity of TOEFL examinees. TOEFL research reports, *Report 60*. Princeton, NJ, USA: Educational Testing Services.
- Kirsch, I., Jamieson, J., Taylor, C. and Eignor, D. (1998). Computer familiarity among TOEFL examinees. TOEFL research reports. *Report 59*. Princeton, NJ, USA: Educational Testing Services.
- Messick, S. (1989). *Validity in educational measurement* (3rd ed.). (ed) Robert Linn. London: Collier Macmillan Publishers.
- Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effect for passage-based tests. *Journal of Technology, Learning and Assessment* Vol. 2 (6) pp. 1-44.
- Russell, & Haney, W. (1997). *Testing writing on computers: An experiment comparing students performance on tests conducted via computer and via paper-and-pencil*. Educational Policy Analysis Archive, vol. 5 (3).
- Russell, M. (1999). *Testing on computers: A follow-up study comparing performance on computer and on paper*. Educational Policy Analysis Archive, vol. 7 (20).
- Sawaki, Y., (2001). *Comparability of conventional and computerized tests of reading in a second language*. *Language Learning & Technology*, vol. 5 (2) pp. 38-59.
- Taylor, C., Jamieson, J., Eignor, D. and Kirsch, I.(1998). The relationship between computer familiarity and performance on computer-based TOEFL test tasks. TOEFL research reports. *Report 61*. Princeton, NJ, USA: Educational Testing Services.

- Taylor, C., Kirsch, I., Eignor, D. and Jamieson, J.(1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning*, vol. 49 (2). pp. 219-274.
- Whitworth, B. (2001). *Equivalency of paper-and-pencil tests and computer-administered tests*. Unpublished dissertation, University of North Texas.

# **DEVELOPING ASSESSMENT FOR LEARNING THROUGH E-TIVITIES**

**Alejandro Armellini, Sylvia Jones  
and Gilly Salmon**



# Developing Assessment for Learning Through e-Tivities

Alejandro Armellini, Sylvia Jones and Gilly Salmon.  
Beyond Distance Research Alliance, University of Leicester, UK.  
[Alejandro.Armellini@le.ac.uk](mailto:Alejandro.Armellini@le.ac.uk). Tel. 0116 252 2768.

## Introduction

Online learning activities or *e-tivities* (Salmon, 2002), in their various guises, can provide for the development of socialisation, teaching, learning, and assessment for students in Higher Education. There is substantial evidence to suggest that learners' engagement with online contributory work correlates with the strength of the link between those activities and assessment: the stronger the link, the higher the engagement (Dweck, 1999; Taras, 2001; Bernardo et al, 2004; Rovai, 2004). The Adelie project aims to embed good practice in re-design for e-learning, build capacity within the institution and enhance the learner experience. This paper focuses on assessment *for* learning (Black et al, 2004), as opposed to assessment of learning through e-tivities. We present a framework for linking e-tivities to assessment.

## Background

Adelie is a one-year Higher Education Academy-funded Pathfinder project aimed at embedding sustainable and pedagogically sound e-learning practice across the University of Leicester, with a focus on re-designing to meet the e-learning and assessment needs of specific academic departments. In doing so, it builds capacity among University of Leicester staff. By bringing together pedagogy, subject knowledge and e-learning design, Adelie researches change occurring as a result of the normalisation of sustainable e-learning practice at three different levels: institutional, teaching practice and learner experiences. This paper presents the main findings of this project in the area of improving assessment *for* learning through e-tivities.

The Adelie Project attracts small teams of academics working together on a new online course, or on an existing course that will incorporate online components. Part of the research team's work within Adelie involves running two-day discipline-specific workshops called *Carpe Diem*. During *Carpe Diem*, teams are invited to reflect on appropriate assessment practices as they design online elements for their course. Among other activities, participants develop e-tivities that align teaching practice, learner engagement, formative assessment and summative assessment. By the end of day two, they have a set of relevant e-tivities running on Leicester's virtual learning environment (Blackboard).

For the purposes of understanding the process of embedding and adjusting our interventions to encourage and disseminate good e-learning design across the university, data is collected at various stages of *Adelie*. Interviews are conducted before and after *Carpe Diem*. Observations during the workshops are carried out. The data is analysed using QSR N6 (Nudist). The e-tivities produced during and after the workshops are also analysed, in particular their links to assessment. By April 2007, 13 *Carpe Diems* across 11 disciplines had been run, involving 70 academics who developed in excess of 50 e-tivities.

## Findings

E-tivities designed by course teams during *Carpe Diem* workshops are varied in terms of purpose, clarity, design, length, demands on the learner and use of technology. Some are clearly written for formative purposes, while others are of a summative nature. The use of interactive discussion boards is central to some e-tivities but marginal in others. Fit for purpose e-tivities, whether in a blended or distance learning course, identify core or threshold concepts (Meyer and Land, 2006) and provide a scaffold for the appropriation of these. They build interactivity, collaboration and opportunities for independent and inter-dependent learning into their design.

Assessment is a key catalyst for change during *Carpe Diem*. Assessment shapes and constrains course design and the design of e-tivities. At stake in the decisions is tutor time and fears of plagiarism. Some subject teams are reluctant to put resources into e-moderation and the formative assessment of students' work. These participants do not regard online collaboration, peer feedback or self-assessment as relevant (Nicol and Macfarlane-Dick, 2006). Other subject teams are suspicious of collaborative e-tivities which, they claim, cannot be part of the assessment because of what they regard to be an opportunity for plagiarism.




Subject teams planning and designing e-tivities for assessment followed the models presented in Table 1. These models are grounded in the data collected through observation of *Carpe Diem* activity and in the analysis of the e-tivities that the teams designed. They show four typical responses to the problems of designing for learning and assessment.

	<b>Links between e-tivities &amp; assessment</b>	<b>Rationale</b>	<b>Tutors' actions</b>
1	Output of e-tivities is (part of) the assessment.	All e-tivities designed to be assessed and may replace essay.	Assess after submission.
2	Two sets of e-tivities: compulsory and optional.	The former to carry a proportion of grade and may replace essay, the latter not formally graded.	Assess compulsory e-tivities. Some e-moderation and monitoring needed.
3	E-tivities are optional, but their output clearly builds towards an assessed assignment.	E-tivities designed and sequenced to align the development of ideas and content with the requirements of a subsequent assessed assignment.	Formative feedback as part of sustained e-moderation is paramount.
4	E-tivities are optional (not assessed).	Keen students given opportunity to learn more.	E-moderation optional but key to maximise learning opportunities and do justice to contributions.

**Table 1: Links between e-tivities and assessment**

The third model shown in Table 1 is perhaps the most interesting and the hardest to design for. It involves a sequence of structured e-tivities, whose design is conceptually aligned with the rubric of the assignment. It also requires a significant e-moderation component. Assessment can be a lever for effective learning if appropriate scaffolding is provided in the form of well-designed e-tivities and good e-moderating practice.

The following e-tivity, adapted from the work of a course team during *Carpe Diem*, is part of the sequence that builds towards the final assessed assignment and provides an example of assessment for learning through e-tivities. It was designed as part of a postgraduate course in Occupational Psychology and is intended to teach a core element in the course, performance assessment. The task structures and scaffolds key aspects of the assessed assignment and thus illustrates model 3 of Table 1.

	<p><b>E-tivity 3a: Is Performance Appraisal Working?</b></p> <p><i>This e-tivity helps you plan the content of the report you are required to submit for assessment. It is NOT the assignment itself. It is designed to help you complete the assignment.</i></p> <p>You have been given privileged access to one document and two audio recordings. All parties have given their consent for you to see and use this information, which will help you understand some of the issues that you could include in your report.</p> <p>(1) Document: <a href="#">Job, performance and statistics information.doc</a></p> <p>(2) Interviewer Training Audio File (55 seconds).</p>  <p>(3) Audio File of a "typical" performance appraisal for In-Branch Customer Services Staff (2 minutes and 34 seconds).</p> 
<b>Purpose</b>	To identify and elaborate on three key issues on performance appraisal.
<b>Task</b>	Identify 3 major issues that arise when you have listened to and read these resources. In no more than 150 words explain why you have chosen these 3 issues. <a href="#">Post your message to the discussion group</a> by <b>Friday 2nd March 2007</b> .
<b>Respond</b>	By the <b>Friday 9th March 2007</b> return to the forum and <a href="#">elaborate on one or more of your fellow participants' posts</a> , responding to their arguments.

## Conclusion

E-tivities and assessment may be effectively integrated into course design following any of the models shown in Table 2, for both on-campus courses with online components and distance ones. While there are no *a priori* right or wrong options, we associate the notion of *assessment for learning* with the third model shown in Table 2. If learners have addressed the sequence of e-tivities responsibly and strategically, they will have a large proportion of their final assignment conceptually written by the end of the course. Being explicit about the link between e-tivities and the final assignment, providing timely and adequate feedback through effective e-moderating techniques (Salmon, 2003) will generate focused, meaningful and purposeful contributions. These will, in turn, lead to improved assessment results and a more positive learning experience.



## References

- Bernardo, V; Parente Ramos, M; Plapler, H; Poli de Figueiredo, L; Nader, H; Silva Anção, M; von Dietrich, C and Sigulema, D (2004) Web-based learning in undergraduate medical education: development and assessment of an online course on experimental surgery. **International Journal of Medical Informatics**, 73, 731-742.
- Black, P; Harrison, C; Lee, C; Marshall, B and Wiliam, D (2004) **Working Inside the Black Box: Assessment for Learning in the Classroom**. London: NferNelson.
- Dweck, C (1999) **Self-theories: their role in motivation, personality and development**. Philadelphia, PA: Psychology Press.
- Meyer, J and Land, R (eds) (2006) **Overcoming barriers to student understanding: threshold concepts and troublesome knowledge**. London and New York: Routledge.
- Nicol, D and Macfarlane-Dick, D (2006) Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. **Studies in Higher Education** 34(1) 199-218.
- Rovai, A (2004) A constructivist approach to online college learning. **Internet and Higher Education**, 7, 79-93.
- Salmon, G (2002) **E-tivities. The key to online learning**. London: Kogan Page.
- Salmon, G (2003) **E-moderating. The key to teaching & learning online** (2<sup>nd</sup> edition). London: Taylor and Francis.
- Taras, M (2001) The use of tutor feedback and student self-assessment in summative assessment tasks: towards transparency for students and for tutors. **Assessment & Evaluation in Higher Education** 26(6) 605-614.



# **ACCESSIBILITY IN E-ASSESSMENT**

**Simon Ball**



# **Accessibility in E-Assessment**

Simon Ball  
Techdis

## **Abstract**

E-Assessment offers many opportunities to broaden the range of tools at the assessor's disposal and thereby improve the overall accessibility of the assessment experience. In 2006 TechDis commissioned a report, produced by Edexcel, on the state of guidance on accessibility at the various stages of the assessment process - question design, construction of delivery software and so on. The findings from this report will be briefly presented, and discussion with participants will be held to ascertain priority areas for the development of guidance for the sector.

## **Introduction**

In 2006 the JISC TechDis service commissioned from Edexcel via a tendering process the production of a document entitled "Accessibility in e-Assessment Guidelines", following consultation with the E-Assessment Group (membership listed on page 4 of the report), to examine the state of guidance for accessibility in e-assessment in the UK.

The purpose of the discussion session at the CAA Conference is to stimulate debate of the issues highlighted in the report, with the hope that some of the key stakeholders in this area might commission or fund further work to formulate the guidance that the report has established is required in the sector. While TechDis is willing, as part of its regular programme of activity, to coordinate this work in order to improve the provision of inclusive assessment in the UK, current funding models do not extend to the commissioning of a piece of work of this magnitude.

Your comments regarding the content of the report and its findings would be most welcome, either during the session itself, or via email at any time.

## **Report Ethos**

This work is based upon the following convictions:

- E-assessment should be fundamentally more accessible than paper based assessment.
- Accessibility in the widest sense is a fundamental quality criterion for assessment and should be considered through the assessment lifecycle.

- Good practice in accessible design will help future-proof assessments.
- Accessibility design should be equally applicable to all assessments (the start of the process being consideration whether there is a reason why a particular assessment should not be made accessible for any reason!).
- Accessibility design should be a changing approach as technology and experience develop. Hence it is a holistic attitude and approach rather than compliance to a rigid checklist.
- No claims are made for the relative costs and benefits of upstream consideration of accessibility compared to post delivery modifications, but organisations are legally and morally obliged to demonstrate that their approach includes all reasonable steps.

## **Principles for accessible E-Assessment**

There are 4 key principles which should be applied to define e-assessment developers and providers working practices:

### *Principle of Anticipation*

The developer should anticipate the variety of accessibility needs that may occur and seek to design in solutions to minimise the through life cost of accessibility.

### *Principle of Reasonable Accommodation*

One of the factors in assessing what is a reasonable adjustment is the overall resource available to the organisation. For example the DRC guidance recognises that capital budgets limit the timescale within which an organisation's existing facilities may be adapted, so it may be acceptable to only convert one building for accessibility if multiple training facilities are available. Therefore although an assessment provider may identify many steps that could be taken to improve accessibility, they may make reasonable judgements as to what is achievable in a given timescale.

### *Principle of Ongoing Technology Change*

It is recognised in the DDA that the continuous advances in technology means that over time new methods of providing accessibility will become available in either absolute or justifiable expense terms. Therefore there is a requirement on organisations to have a process of continuous review of their approach to accessibility.

### *Principle of Corporate Responsibility*

The responsibility for complying with the DDA rests with the organisation and hence its senior management. To ensure that all the individuals in an organisation make consistent efforts to comply with the Act, an organisation's management should ensure that there is a clear accessibility / anti-discrimination policy, training to ensure compliance and a monitoring / review

process to check that the policy and training are being followed and are being successful in achieving compliance with the DDA.

It is the authors' belief that there is a legal and moral requirement upon Awarding Bodies and related organisations to have a demonstrable commitment to each of the 4 principles above. This must be demonstrated by the most senior management on down through the organisation. As each Awarding Body is in a unique position regarding adoption of e-assessment and the type and maturity of technology involved, each organisation must develop its own response to these principles.

### **Practical steps towards accessible E-Assessment**

To follow the key principles above, there are a number of practical steps a supplier of e-assessment products can take:

1. Develop/amend internal processes and procedures to reflect the accessibility "good practices" identified in the various accessibility documents and websites identified in the Codes of Practice.
2. Implement training, tools and product auditing to ensure that compliance with the processes and procedures is achieved.
3. Develop a "technology roadmap" for accessibility and produce a plan with resources and timescales to implement it. This is likely to include identifying a list of preferred accessibility tools and working with suppliers and customers to ensure their technical support and use. It may also include the development of tools to assist the processes and procedures from item 1.
4. Implement an ongoing review of the success and applicability of the above 3 steps on an annual basis.

### **E-assessment development process**

Different organisations will have unique development processes, which will vary dependant upon factors such as whether technical resources are in-house or subcontracted and the e-assessment is targeted to general or professional qualifications.

Consideration should be given to accessibility and usability issues at each stage. Organisations should review their working methods and own development processes but may wish to use the suggestions below as an initial model.

### **Test Specification**

In relation to qualifications the DDA makes a key distinction between an awarding body's duty to make reasonable adjustments to the assessment process and its right not to adjust the competence standards inherent in a

qualification. The specification must (among other things) therefore address two key issues:

- Complete clarity on the competence standards underpinning a qualification and which of these are mandatory – hence this establishes at the outset what justifications may exist for providing a non-accessible assessment,
- Definition of whether the competence standards require testing via e-assessment. If this is not the case, alternative equivalent means of assessment (e.g. a paper assessment) may be considered as one method of ensuring accessibility.

### **Development Team**

A test development team for a major assessment is likely to be distributed, often drawn from more than one organisation (especially where on-screen content or delivery technology is subcontracted). It can typically include;

- Principal assessor
- Test specification author
- Test author
- Content producer
- Delivery platform provider
- Accessibility specialist

The consultation showed that most organisations feel satisfied that their internal processes for ensuring accessibility are robust and well practised and that upcoming changes to legislation are anticipated and will be addressed. This included most organisations having specialists to develop accessible versions of existing assessments. The most significant change for e-assessment (and arguably paper assessments!) is that these specialists should be involved at the initial stages onwards. A potential weakness is in multiple agency/organisation development where understanding of practices, capabilities and techniques may not be shared. It is recommended that once a team is identified, the responsibilities of the parties are identified in writing and that where a team and / or the technology to be used are new, a joint capabilities training session is held. This ensures that:

- the specification takes full advantage of the capabilities (e.g. multimedia), whilst recognising any limitations (e.g. security lock-down limiting assistive software),
- the author and content producer agree on all relevant information required to define an item,
- the items are authored to take advantage of innovative capability,
- required developments to the content delivery platform are identified early.



The consultation suggests that this is currently an ad-hoc and sporadic process.

## **Test Requirement Document**

The test requirement document must capture the preferred assessment method and the requirements for accessibility. In particular, where the required competence standards indicate that the assessment cannot be made accessible to certain disabilities, this should be stated. Where accessibility is required consideration should be made at this stage whether it is through:

- the application of technology (assistive software and aids),
- other supportive measures (e.g. a reader or scribe),
- alternative means of assessment (i.e. a practical rather than simulated test).

Definition at this stage means that the requirements on the various members of the development team are clearly stated and development funds are spent on the identified areas of accessibility.

Where the specification calls for simulation care must be taken on two fronts:

- If the actual implementation is emulation then existing accessibility approaches may not work (e.g. an emulated software package in ICT testing may not support all accessibility functions available in the full package).
- The simulated environment may not be rich enough to reflect how individuals work in practice (e.g. a simulation cannot replicate the sense of touch to explore shape and texture).

Where simulation is specified the three options above for alternative assessment must be carefully considered.

In the specification, standards should be invoked with care. Invoking standards does not confer accessibility or a given level of quality. For example, a test item may be IMS QTI compliant, but that does not define how it will be displayed on-screen and hence how usable it is!

## **Write test**

The author should write the test with the specified assessment method and technical capability in mind. If the technology to be used is new to the author, they should be trained by the technology provider to understand the capability of the content and delivery system, and the information required by the content producer.

The principal difference to authoring the paper test is that a deeper level of description and detail is required to fully describe what is being tested, how it

is to be tested, and how various elements of the technology should handle the test data:

- Where a qualification is only partially accessible due to the underpinning competencies required, a statement should be provided if a particular question is not to be accessible in certain aspects (for example an vocational test of electrical engineering may test that the candidate knows the wiring colour code, which will be fully accessible, whereas a practical test of recognition of colours and hence correct wiring cannot be made accessible to colour blind candidates through colour labelling).
- Stating the competencies being tested in an item ensures that the content producer does not provide unfair assistance through accessibility measures – for example where a written comprehension should not have a voice-over.
- To reflect simulation or multimedia approaches, a storyboard may be most applicable.
- Where the data is available to the candidate in multiple forms (e.g. written, graphics, alt text and sound effects) each should be specified if critical to the equivalence of different methods of access.

This guidance is in addition to the general guidance produced by the regulators on issues such as use of appropriate language, representing diversity, avoidance of bias etc, which applies equally on-screen as to paper assessments.

## **Write Mark scheme**

Current JCQ guidance published on 6th September 2005 in response to the pending extension of the DDA to general qualifications, is that all qualifications should be allocated on the same mark scheme without exemptions and a subsequent certificate indication. This means that the mark scheme should be written without consideration of the specified level of accessibility – that is where certain skills cannot be demonstrated by a person with a disability, a mark cannot be provided that excludes that skill (an indicated award).

This removal of any consideration is a rather perverse (and unintended) effect of the equality legislation and is likely to come under significant scrutiny and possible revision. One alternative is to ensure that qualifications are designed on a unitised basis where the units are designed such that one or more units may contain all elements relating to a competence that may by definition be inaccessible to some candidates.

There is a particular area of interest and uncertainty here with item bank based tests. Ultimately, if questions with varying degrees of accessibility can be argued to be an equivalent test of a competency, an Awarding Body may choose to create tests ‘on the fly’ from an item bank using accessibility criteria

as one of the elements of the selection algorithm. This will only be possible if there is a rigorous mark scheme which ensures that the algorithm selects a fully representative test for candidates selecting an accessible option.

## **Test QA Process**

The first stage of QA assessment is to check that the test requirement fully reflects the test specification and to ensure that the test items are satisfactory in terms of validity, reliability and accessibility in its widest sense. The standard processes used for paper examinations are well practised and understood, and are a first stage for the on-screen QA process.

The requirement for on-screen is an extension of this process in that the QA process must also check that:

- The author has specified what accessibility options are not applicable due to competency requirements,
- Allowable accessibility options are fully specified,
- That the accessible version (i.e. voice-over, alt-text etc) is comparable for difficulty.

A key difference for on-screen authoring, as with software publishing in general, is that many pieces of independent code, each of which has a unique revision state, may be brought together to make a complete assessment. The awarding body with ultimate responsibility for the assessment must ensure that the organisation authoring this code has a suitable robust configuration control system in place which enables the tracking of each piece of code, including traceability of review comments and subsequent modifications. Each subsequent release of the assessment should then have a revision designation which enables the revision state of each element to be determined. This is not a unique issue for accessibility design, but is a necessary step to ensure that changes requested as a result of accessibility checks are tracked and properly implemented.

## **Mark Scheme QA Process**

The mark scheme QA process for on-screen assessment is essentially the same as for a paper assessment, however there are two key checks that should be undertaken:

1. If an on-screen assessment is to be marked automatically the mark scheme must define acceptable boundaries of data entry (for example are typographical errors to be penalised), so that a suitable marking algorithm may be developed.
2. The interaction of the mark scheme and the screen based interaction should be considered, such that the assessment does not become an inadvertent test of dexterity / motor skills through the allocation of marks for a solution that is not keyboard or switch navigable.

## **On-screen authoring**

Professional on-screen authoring organisations should be expected to have 'style manuals' which provide their authors with guidelines on how to develop items following good practice for both accessibility and usability. Key issues that should be addressed are:

- Ensuring there is good communication with the author should clarification of the specification or acceptability of approach be required.
- Train on-screen authors to recognise the impact their authoring decisions may make on item difficulty and comparability.

## **Marking algorithm implementation**

Following on from the consideration of the initial mark scheme design, a key aspect for any on-screen marking algorithm is to implement the specified level of robustness to candidate entries. Whilst straight-forwards for multiple-choice based knowledge tests, this may include such innovations as neuro-linguistic programming for the assessment of free text entries.

Also it is important that where the output rather than process is being assessed, the algorithm does indeed check output and does not use process as a proxy – for example some ICT tests mark 'process' and therefore fail to give marks when users use less common working methods for accessibility reasons.

## **QA draft assessment**

Each Awarding Body will develop their own quality assurance process in agreement with their technology provider (third party or in-house), which should explicitly checks that accessibility features are included and operating as specified, and that the validity and difficulty of the assessment is comparable for each alternative method of access.

## **User acceptance testing**

Typically in accessibility much consideration is given to a purely technical review of accessibility. However the core of the exercise is to produce e-assessments that users find both meaningful and manageable. The only way to ensure this is through user trialling.

User trialling is a challenging and time consuming business which becomes much more so if attempts are made to trial with particular user groups, such as users of particular assistive devices and those with a particular disability. This should be addressed through a layered testing approach, with the e-assessment delivery engine, generic content (i.e. questions types) and specific content (i.e. actual questions) having different assessment regimes.

For example whilst a delivery engine and generic question type may undergo testing for navigation using particular assistive technologies, once proven, this need not be repeated for each subsequent use of that question type.

Each Awarding Body should develop their own system of user testing and be able to demonstrate that there is a robust system of recording user comments (which will include centre staff), feeding back comments to authors and content producers, and tracing modifications to the assessment to maintain quality.

### **Sign-off assessment**

The assessment sign-off indicates that the level of checks considered reasonable within the awarding body's own QA process has been passed. The major issue for accessibility is that the majority of real accessibility testing will happen in the field, post sign-off whereas the sign-off and release process should allow for the collection of field usage data and the subsequent update of an assessment and feedback to authors and content producers.

### **Operational Roll-out**

Operational roll-out should comprise two distinct phases:

- Initial implementation in centres,
- Ongoing feedback and improvement.

The consultation suggests that the former is an effective process with centres. There are existing standards such as BS7988 which provide information on the generic standards that ICT test facilities must follow. Awarding Bodies and their technology partners further have guidelines on particular issues such as equipment specifications, staff training, required roles, defined points of contact, escalation routes etc, which this document does not seek to replicate.

However the consultation does indicate that the main area for potential improvement is the on-going feedback and improvement. The delegation of responsibility for applying adjustments to centres appears to have had the effect of limiting the flow of information on accessibility issues from the centres to the awarding bodies and their technology providers. It is not clear whether the low volume of requests for accessibility support from centres to Awarding Bodies reflects a high level of self capability or an indication that candidates are either being steered away from on-screen assessments by centres or choosing themselves not to enter for on-screen assessments.

If it is the former, then there is potentially a large body of evidence and skill on how to integrate accessibility technology, which could be collected and made available on a wider basis. If it is the latter, then there is a need to improve the communication.

The ideal approach is that centres should have a defined point of contact for accessibility issues and be encouraged to provide user feedback both on what does work and proves popular and what accessibility aids have been tried but failed to interoperate. This can then be used to create a knowledgebase to inform future developments and support other centres.

### **Cost – benefit analysis**

This report purposely avoids making statements as to the relative costs of alternative approaches, or what costs may be determined ‘reasonable’ in legal rulings under the DDA. However what is clearly not good practice, and demonstrates a poor culture of accessibility and usability is proceeding with a development of on-screen assessment and at a late stage of the process, calculating the cost of ‘adding-in’ accessibility features, comparing the cost with the ‘expected’ number of users (particularly if based upon past data on requests for modifications) and using this as a justification not to adopt accessibility options on the basis of a ‘not reasonable costs’ defence. Such an approach is poor on several counts:

- It perpetuates existing design approaches and stifles innovation,
- It assumes the past, with all the barriers to accessibility, is a good indication of how many people will aspire to qualifications in the future,
- It ascribes no value to the benefit of good usability to the wider population.

### **Consultation Findings**

During the development of this document, the authors consulted with a number of organisations including government agencies, awarding bodies and technology providers. In addition to what has been described above, the major points or issues are recorded below.

1. All parties consulted on e-assessment believed that there was a good level of understanding on the need to comply with the DDA, and that there was much generic (generally web-derived) assistance on on-screen accessibility techniques. There is an issue that knowledge of how to apply the legislation and case law to confirm the principles of application are both still evolving. The regular and wide sharing of such information, as it becomes available, would be most useful.
2. A possible means of sharing both best practice and emerging guidance and case law would be through an online forum for awarding bodies and technology providers. TechDis already provides considerable useful resources and an online forum could be created as an addition to that support.

3. 'Reasonable cost' justifications for not adopting measures to improve accessibility and usability typically do not allocate any 'benefit' value to the usability element of the cost-benefit calculation despite diverse surveys from an assessment of Tesco.com to Microsoft usability surveys indicating broad benefits from adopting good accessibility practice.
4. Awarding bodies are not technology specialists and interoperability issues (between assessment platforms, assistive software and technology) are continually changing as technology advances. Specialist centres are reportedly well placed to support individual students but there is little evidence of feedback into the platform or assessment design process. Also technology providers undertake ad-hoc testing for interoperability, but there is no formalised recording of interoperability or sharing of data. There is interest and potentially significant benefit in having a centralised organisation that has access to assistive technology and trained users that can facilitate compatibility and usability testing with trained users. This could provide a coherent UK lobby voice to major software suppliers, as well as a central point of contact for learners, test centres and technology suppliers for information and support.
5. The issue of language as an enabler was raised in consultation – the assertion being that it is typically an un-stated criteria. This is particularly the case in vocational qualifications and is significant for on-screen testing where many assistive aids are potentially available such as voice-overs, clear iconography, on-line dictionary, spell checker and thesaurus. There is an argument that the required level of language should be explicit, and the level of acceptable support be defined to avoid a disparity between an on-screen test and the 'equivalent' alternative practical or written test.
6. The point above may be linked to the apparent improvement in test results by moving from a paper test to an 'equivalent' on-screen test. Other reasons have been postulated such as a reduction in exam stress through a non-threatening environment and reduced distractions through presentation of a single question on-screen at a time. It is clear that there is a fine line to be walked between providing comparability and accessibility / usability. This area whilst not directly related to accessibility and usability is clearly important and would benefit from further research.
7. The assertion was made that integration between authors and content producers and design for accessibility is better in learning content and assessment for learning than in accredited qualifications – possibly through considerations of security and equivalence and possibly through custom and practice of existing development teams. There may be some benefit in looking to non-accredited test and content developers for good practice.

8. There are two wider inclusion issues for centres and learning providers to consider; how to encourage wider participation in learning and assessment and what the implications to moves towards e-learning and e-assessment means for those with no access or poor skills in ICT.
9. There are many standards relating to the technical aspects on-screen assessments and accessibility of web sites / onscreen material. However there are variations on how close to market they are, how they relate to functional specifications and whether there are contradiction between standards or significant gaps left to 'interpretation'. There is also not a known standard for accessibility testing of assessments – most work in this area just relates to web design and therefore misses some significant aspects of assessment design such as security and reliability. The area of standards has not been significantly covered here and would merit further consideration.
10. The issues raised in consultation are primarily concerned with timed assessments. E-portfolios are used for accredited qualifications, but as this is typically output based (e.g. DiDA), reflecting a candidates normal working practices, there is considerable scope for learning providers to take individual measures for accessibility and hence e-portfolios in a general sense are not considered problematic. However the recent e-portfolio report for Becta highlighted that where an e-portfolio platform is mandated, many are poor on issues of accessibility, usability and inclusion!
11. The increasing use of technology reflects the wider world in which learners operate and the drive by awarding bodies to find a competitive advantage. Respondents to the consultation were generally satisfied with a 'light touch' regulatory approach, where Centres, Awarding Bodies and their technology partners put forward proposed approaches and their justification for using an approach, rather than asking the regulator to make sweeping rulings in advance of developments for example, in the development of innovative item banks, the exact rules for an algorithm to select questions and allow time based upon disability should be open to development and proposal rather than being prescribed.
12. As the current system delegates the responsibility for providing access to the test centres, there is little or no information collected or collated by the Awarding Bodies. This means that there is little centralised information on the level of use of various assistive technologies and whether improvements in design result in an increasing take-up of e-assessments by candidates with disabilities.



## **Conclusion**

This report raises some useful, interesting, and potentially contentious issues. The aim of presenting this report to the audience of the CAA conference is to stimulate debate and obtain feedback on the most appropriate way forward for TechDis in this area.



# **THE VERIFICATION OF IDENTITY IN ONLINE ASSESSMENT: A COMPARISON OF METHODS**

**Dr Trevor Barker  
Stella Lee**



# **The Verification of Identity in Online Tests: A Comparison of Methods**

Dr Trevor Barker, Department of Computer Science  
University of Hertfordshire  
AL10 (AB

Stella Lee, Department of Criminal Justice  
University of Hertfordshire  
AL10 (AB

## **Abstract**

Having an online assessment system that is secure, effective and efficient is a major problem for distance learning providers. At the University of Hertfordshire in the UK, the Criminal Justice Team (within the Social, Community and Health Studies) and the Department of Computer Science are two departments that deliver significant e-learning programmes.

The increased use of online learning systems in education today has in most cases been a positive influence on the learning experience, for learners and teachers alike. The University of Hertfordshire's Managed Learning Environment (MLE), StudyNet is used within our university in a blended framework for learners both on and off campus since 2001. In 2003/4, 80% of staff and students were using StudyNet regularly with 3.62 million logins. In 2004/5, this figure had grown to over 95% of staff and students (4.85 million logins) including 51% of logins from locations off-campus. The use of technology in the assessment process is a logical progression following from the rapid advance of learning technologies.

Well designed assessment is linked with major gains in student attainment and reinforces good curriculum practice. (Ridgway, McCusker, & Peard, 2006). This is particularly important for distance learning courses as there are very little built-in interaction between instructors and students. It is interesting to note that in 2006, for the first time, online assessment and on-demand testing were used in some subject areas by the Scottish Qualifications Authority. Assessment was carried out in school centres, with learners having to attend. Within the UK, some 5,000 on-screen tests are now being taken each week. (SQA, 2005) Other online assessment such as the (insert some examples here) are in progress and are having good results and feedbacks. Issues related specific requirements for online assessment, including security have been given by Roan (2003).

The Department of Computer Science has several hundred remote online learners in locations as far as Trinidad, India and China as well as in the United Kingdom. This number is increasing year by year and is likely to do so in the future. Providing reliable, valid and fair assessment for these learners is difficult and expensive. It is essential that all concerned in the process are confident in the validity of our assessment processes as they are related to the perception of the quality of our qualifications. For this reason, remote learners are required to attend examinations and tests at the University in the United Kingdom or at assessment centres in other countries, in order that tests are properly supervised.

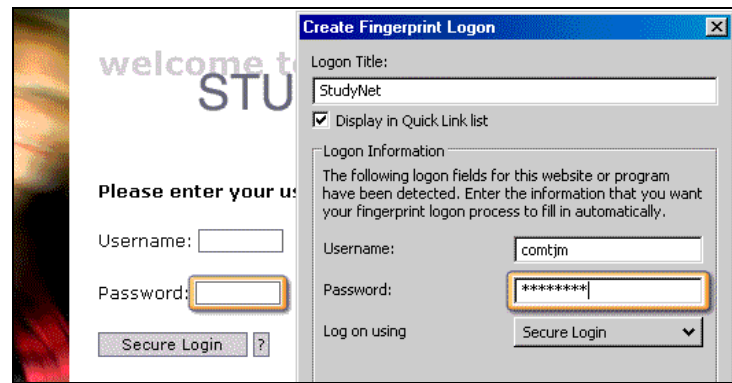
The research presented in this paper is the result of a collaboration between the two departments seeking a reliable and efficient way of providing an online assessment system capable validating the identity of students remotely

In the first stage of our research, the use of fingerprint recognition was tested. Finding from these studies suggested that although this system proved to be useful and reliable,, they were difficult to implement and manages remotely at some locations. Tests with the Microsoft fingerprint reader showed that it was relatively simple to install, important for remote learners. Registration of fingerprints was through a wizard. The finger print registration wizard allows you to register fingerprints by selecting the preferred hand and finger and then following the prompts to scan. Each finger must be successfully scanned four times and you can register up to 10 fingers. You can register other fingers at any time by using a registered finger to access the registration wizard.



**Figure 1: Fingerprint recognition system.**

Creating a fingerprint login is simply a matter of going to the desired login page and scanning any registered finger. You then have the option to associate a username and password with the login page. Once complete, any registered finger will be able to use the login.

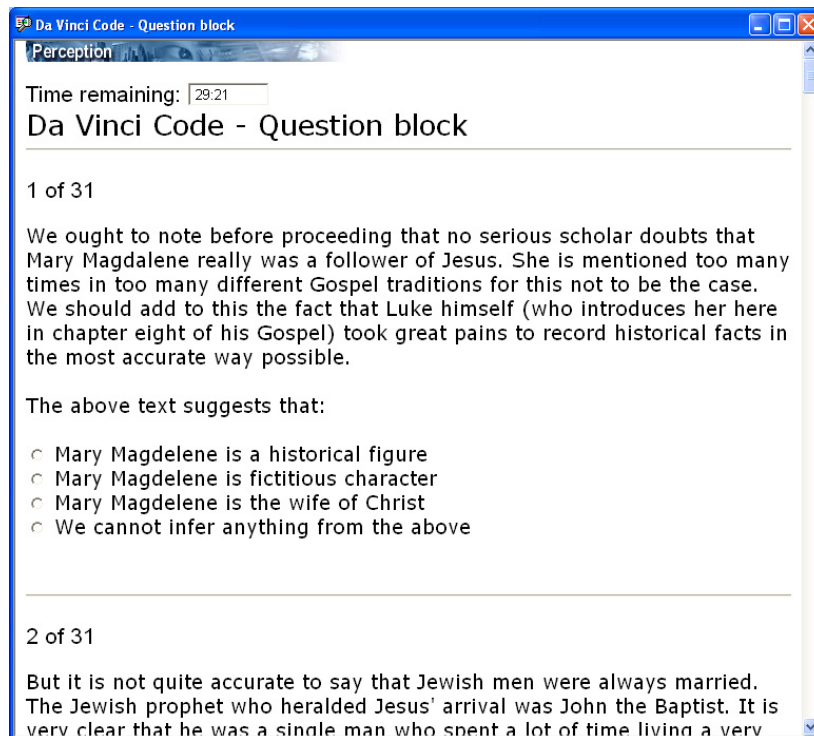


**Figure 2: Fingerprint logon screen**

Any registered finger (user) will be able to delete other registered fingers and disassociate usernames and passwords from login pages.

Several limitations with the system were noted. The system assumes that the user is always the same person. This is because it is limited to 10 scans per user login (although there is nothing to stop you from registering 10 fingers from 10 different people). There is no way to associate a single registered finger with a login page. All registered fingers are associated. The Microsoft fingerprint reader does not work with browsers other than IE. Other scanners on the market will work. The management of such systems at remote sites would be difficult to control and without a significant investment in time and effort.

The next stage of our study was to develop and implement an online identity verification system using video conferencing, a database of user verification information and online chat. Video has been used in several contexts in supporting online education. It has been used for a range of activities such as interviewing for PhD candidates from overseas (Basiel, 2006), team meetings, project supervisions, and focus group sessions. Mostly, it is used for pedagogical, motivational, and content/information delivery. In this study we decided to develop and use an online video streaming system using web cams, integrated with online chat in order to verify the identity of test candidates. The system developed was in two main parts, a management system and a client side application. The system employed readily available Skype video and chat software integrated with the well-known Questionmark Perception assessment software and an online verification database built for the purpose. The test comprised of "reading comprehension" questions loosely based on the book "DaVinci Code". However, no prior knowledge of book was required. The assessment was set up primarily as a reading comprehension exercise and the topic was picked at random. Below (figure 3) is a screenshot of the test used for the two sessions.



**Figure 3: A screenshot of the “DaVinci Code” reading comprehension test**

Two studies were conducted in order to develop and test this system. In the first study, an expert evaluation of the software system was undertaken in order to refine and test the system and evaluation tools.

Comments from the expert evaluators was in general favourable, though potential problems were identified.

“The experience of using the video online test was an interesting one. It was not as intrusive as I had imagined it to be. The small interface makes it easy to launch/operate and carry on with my test at the same time. I did have to pay extra attention (and thus, becomes a bit of a distraction) while I was trying to concentrate on the test questions as I would never know when the video would pop up again. Another issue with using the video is that it was not clear if I were to look up on the web cam to acknowledge the moderator. For the most part, it worked rather smoothly and didn’t cause extra stress or major distraction to my overall testing experience.”

*(Educational Systems analyst)*

“For such a short test [35 minutes] I considered the amount of intrusions to be excessive. Although I had finished about ten minutes early I was still interrupted whilst answering questions on at least two occasions. My other concern was that the Skype interface is somewhat confusing and I was spending too much time trying to work out what button to press. Starting the incoming voice call was fine but then I was forgetting to press the 'Start Video' button to activate the webcam. It



might be useful to include a run-through of what exactly will happen when a call occurs.”

*(Technical specialist)*

In the second study, undergraduate students took an online assessment using the system under controlled laboratory conditions. The session was managed by the same moderator who managed the first pilot study. At the start of the session, each participant started the test and was contacted initially using video by the moderator in order to provide a visual comparison with a photograph held in the database. Once this had been established, subsequent communication was through chat and video requiring responses to questions.

All interactions between the management system and the candidate group were captured and recorded for later analysis. The session was also recorded using a video and audio recorder in order to assess the environment and conduct of the session. It was hoped to ensure in this way that bias was not introduced into the trial and to record any difficulties participants and testers might experience. A user questionnaire was handed out at the end of the session, as we wanted to obtain feedback from users on the test and the verification process. Measures were made of the efficiency of the system, in terms of the speed and reliability of the verification process. The session moderator also provided a reflective log of the session in order to record his experiences of managing the session. In table 1 below, the results of the student questionnaire are presented.

	Not Much		Average		Very much
Experience:	1	2	3	4	5
Using Computer	7%		7%	13%	73%
Using Internet	7%		7%	13%	73%
Assessment Online	20%	7%	27%	27%	20%
Video Conferencing	27%	20%	7%	27%	20%
Online Chat	7%		13%	27%	53%
	Poor				Excellent
Quality of Assessment			7%	67%	20%
	Easy				Difficult
Difficulty of Questions		13%	40%	27%	20%
Starting the Session was	40%	33%	13%		13%
	Disagree				Agree
	7%	13%	20%	27%	33%
Identity checking didn't affect performance		20%	7%	7%	67%
Using video was a problem	60%	27%	7%	7%	
Online chat was a problem	73%	13%	13%		
Often distracted by need to confirm identity	53%	13%	13%	20%	
Difficult to know what to do at times	60%	20%	13%	7%	
Easy to answer questions, despite interruptions	7%	20%	20%	20%	33%
Sometimes missed the video or chat	60%	7%	27%	7%	
Messages were clear and easy to follow			7%	20%	73%
Using video was easy			7%	33%	60%
Using online chat was easy			20%	27%	53%
Easy to remember details I was asked for during the assessment		7%	27%	20%	47%
Too many screen to deal with in the session	80%	13%	7%		
Need to validate identity had a large effect on my performance	53%		20%	20%	7%

**Table 1: Student attitude to online assessment and identity verification**

The results of the student questionnaire shown in table 1 were taken to indicate that the use of the online video and chat identity verification process did not unduly interfere with their experience of the assessment. The range of test scores obtained in the study suggested that the level of the test was adequate to provide a good measure of the quality of the identity system.

Analysis of the interaction data recorded showed that it was possible to make approximately one interaction each minute between the session moderator and the candidates in the 39 minutes the session lasted (26 video and 12 chat sessions in all). This was surprisingly few. Video connections lasted on average about 10 seconds; just time enough for the candidate to respond and the session manager to confirm identity by comparing the video image with the database image. Chat sessions lasted slightly longer, about 15 seconds, due to the need to enter text and also because the request to chat was less noticeable than the request to initiate a video session within the interface.

The most common administration function was searching (33 searches), which took on average 31 seconds for each. Searching involved locating information on candidates held in the database and locating and initiating contact with the candidates on the video and chat system in order to make connections. There were 9 errors in total, occupying on average 84 seconds

each. We considered this to be a large number of errors, approximately one every 4 minutes. The software and hardware was installed on good quality computers in a controlled environment by skilled technical support workers.

Errors mostly related to loss of the video and chat connection due to problems with the software employed for this purpose. In some cases, these were probably due to user error, but mostly they were due to errors of an unspecified nature within the software itself. These errors were resolved by re-starting the software and making connection once more. There were no errors due to loss of network connection per se. In order to correct error, helpers were employed to visit each assessment workstation, locate the reason for the error and re-set the software. This would not be possible in a full scale test with remote candidates. It is likely that suitably trained candidates would be able to detect loss of connection and make the necessary reconnections themselves. There were no errors identified that related to the use or function of the management system or the question delivery software. It would be important in a full scale assessment session to test systems adequately prior to the start of a test session, to train candidates in the use of the software and remedial action and also to have a standby method in case a catastrophic error resulting in total loss of communication with a candidate occurred.

## **Conclusion**

The results of the study were able to identify potential problems and sources of error in the systems employed and also provided potential solutions to them. Learners in general did not report that they were disadvantaged by the video and chat system. A major improvement will be to develop a fully integrated system based on the use of video and chat for the verification of identity with many of the functions fully automated. This will improve the efficiency and reduce error rate. A need to train examinees was also identified. In the next stage of our research, we shall implement improved systems for full-scale testing with remote learners in Trinidad. We are not able to ensure that all learners will have a broadband connection and this will introduce additional challenges.

## References

SQA (2005). Retrieved February 7, 2007, from [http://www.sqa.org.uk/files\\_ccc/SQAsPlansForE-assessment-March2005.pdf](http://www.sqa.org.uk/files_ccc/SQAsPlansForE-assessment-March2005.pdf)

(Ridgway, J., McCusker, S. & Pead, D., 2006, Literature Review of E-assessment, *Futurlab series*, Bristol.

Basiel, A., 2006, e-Learning Issues of Web Based video conferencing: A work Based case Study, *Proceedings of the First Annual Blended Learning Conference*, University of Hertfordshire, June, 2006

Roan, M. 2003, Computerised assessment: changes in marking UK examinations – are we ready yet?, *29th Annual Conference of the International Association Educational Assessment*.  
[www.support/iaea/papers/roan.pdf](http://www.support/iaea/papers/roan.pdf)

# **EXPLORING THE POTENTIAL, LIMITATIONS AND USE OF OBJECTIVE QUESTIONS IN ADVANCED CALCULUS**

**N Baruah and M Greenhow**



# Exploring the Potential, Limitations and Use of Objective Questions in Advanced Calculus

N Baruah and M Greenhow

Department of Mathematical Sciences, Brunel University  
mapgnnb@brunel.ac.uk & mastmmg@brunel.ac.uk

## Abstract

This paper describes our experiences with authoring and trialling questions in advanced calculus topics, namely ordinary differential equations, Laplace transforms and Fourier series. These topics are generally taught at the end of the first year or during the second year of a mathematics or engineering undergraduate degree. We expect that many of the lessons learned here will apply to other conceptually-advanced mathematical and scientific content. Typically, what is significant for such content is that many skills are needed from previous exposure to calculus and algebra, and that paper-based questions at this level tend to be more abstract, holistic and open-ended, requiring the sort of flexibility in marking generally associated with human markers. For objective, and therefore more constrained questions, we do not know what is feasible and whether or not questions on advanced topics will actually test the skills they are designed to test. For example, a student may carry out e.g. a Laplace transform correctly, but make an elementary algebraic mistake near the end; this would be easily recognised by a human marker, but simply marked wrong by any current CAA system which cannot assess the (generally handwritten) intermediate steps in a student's solution. Conversely, any question that can be marked by a CAA system is likely to be structured or scaffolded (e.g. by asking for intermediate steps explicitly) so that the original requirement on the student to devise a solution strategy is lost. This paper explores what can be asked effectively: facility with such questions is a necessary (but not sufficient) condition for students to master more advanced topics, so some sort of blended assessment (with human markers) may still be needed for higher-level skills. We describe the process of authoring higher-level objective and report of the experience of running the questions with our second year cohort, including an analysis of the answer files produced. Our evidence suggests that the assessments were useful to students in establishing a solid foundation of skills, mainly by being encouraged, or even forced, to engage with the extensive feedback screens.

## Background

During the current academic year, online tests in the advanced calculus section of Mathletics were authored and delivered in an extended form of Question Mark Perception. Mathletics is designed to exploit the potential of

computer-aided assessment, especially in formative assessment mode. Our experience over the last 5 years of trials with many hundreds of students indicates that they value the extensive feedback (generally including a fully-worked solution) as a learning resource, as well as for the marks awarded. Moreover, students' learning has been encouraged by the tests building the confidence of first-year undergraduates. The pedagogy of building tests into a module is quite well established, and various trials of the mechanics material have indicated that students move, at least partially, to a deeper approach to study (Gill & Greenhow, 2004). This paper examines whether or not the same claims can be made for assessments covering advanced calculus topics delivered to second year undergraduates.

The underpinning technology of Mathletics, whereby many thousands or millions of question *realisations* are generated by a single question *style* that encodes the algebraic and pedagogic structure of the question, is carried through to the more advanced content described in this paper. We have found that it is extremely helpful in moving students away from simple memorisation towards an understanding of the question's content and solution. The random parameters, possibly constrained according to the question's content (realism of the question and reverse engineering from a desirable solution form), are carried through to all parts of the question so that it realises with:

- dynamic MathML, giving equations in the question and in the (often extensive) solution and other content given as feedback.
- dynamic SVG, giving accurate diagrams, charts and graphs.
- dynamic wording, giving different scenarios, expressed in gender- and ethnically-balanced language.
- dynamic question functionality, such as algorithms that, when run to completion, generate, for example, HTML tables of variable length.

Accessibility (SENDA compliance) has been a key feature of the existing questions. The format of all elements may be chosen by the student and stored as a cookie. A great deal of technical effort has also gone into the writing of functions to underpin the questions. These split into two basic types: functions that return the result of a calculation, e.g. multiplying out two polynomials of arbitrary order, and functions that return display strings e.g. a MathML string to display a table of Laplace transforms or an SVG string to display a graph of a function and a few partial sums of its Fourier series, see figure 3. Exportability of the mathematical content to an ordinary web page or other web-based CAA/CAL systems is another key feature, see Ellis, Greenhow and Hatt (2006).

Before the construction of the questions, the learning levels of the questions were categorised from a pyramidal (rather than hierarchical) version of Bloom's taxonomy (Hatt, J. & Baruah, N. 2006), with the six chambers in three learning levels. The pedagogy of each of the subtopics of Laplace transforms and Fourier series was analysed to specify the tested, and prerequisite,



concepts and skills. Concept maps (Turns et al 2000) were drawn for this purpose. The questions are mostly from the first two levels comprising the *remember*, *understand*, *apply*, *analyse* and *evaluate* chambers. *Create*-level questions were designed for only a few topics, see figure 2.

One of the main objectives of the study was to develop questions at higher-learning levels. Several different question types (multiple-choice, multiple-response, hotline, true-false, numerical input and responsive numeric input) have been utilised. For effective and targeted feedback, mal-rules encapsulating the essence of an incorrect solution method or error, are needed for multiple choice, responsive numeric input and hotline questions. To discover such mal-rules, the answer files of previous elementary calculus CAA tests were analysed and answer scripts of past examinations were examined. From such evidence, and from the works of Orton (1983), Schechter (1994) and Greenhow (1996), an error taxonomy has been developed. Not only is this useful in question design, but it also greatly facilitates the interpretation of students' answer files.

For more advanced topics, the choice of question type needs specific attention from both pedagogic and technical standpoints. Some multiple-response questions were designed to test students' understanding of general mathematical properties, but the form of multi-choice questions may be ineffective due to guessing. To overcome this, new four-optioned yes/no and true/false question types have been designed for testing identification of general properties and theorems, see figure 1. Such questions are scored dichotomously to reduce drastically the probability of rewarding guessing. Whilst the question in figure 1 is quite static (in that other realisations will look very similar) other versions are made more dynamic by replacing the unspecified general functions by particular randomised functions.

The hotline and responsive numeric input questions were extended in some of the topics by recording the students' certainty in their answers, along the lines given by Gardner and Gahan (2003), but without negative marking.

As some of the problems solvable by the Laplace transforms naturally require the inclusion of the diagrams, dynamic drawing objects have been developed using Scalable Vector Graphics (SVG). These build elementary drawing objects like lines, rectangles and ellipses, to form new objects such as graphs. Such types of questions may be helpful in the presentation of the question stem, say by the inclusion of circuit diagram, or in the feedback, see figure 3.

An example of a question at the higher *create* level is shown in figure 2 where students need to obtain the limits of integration by correctly interpreting the diagram. Thus students are being tested on concept of periodicity. In the feedback, the general form of the Fourier series is written before being applied to this particular question; this exposes students to the underlying concepts (deep learning) as well as purely procedural skills (surface learning). The feedback is reinforced by providing a graph, plotted using a high-level function due to Ellis (2006).

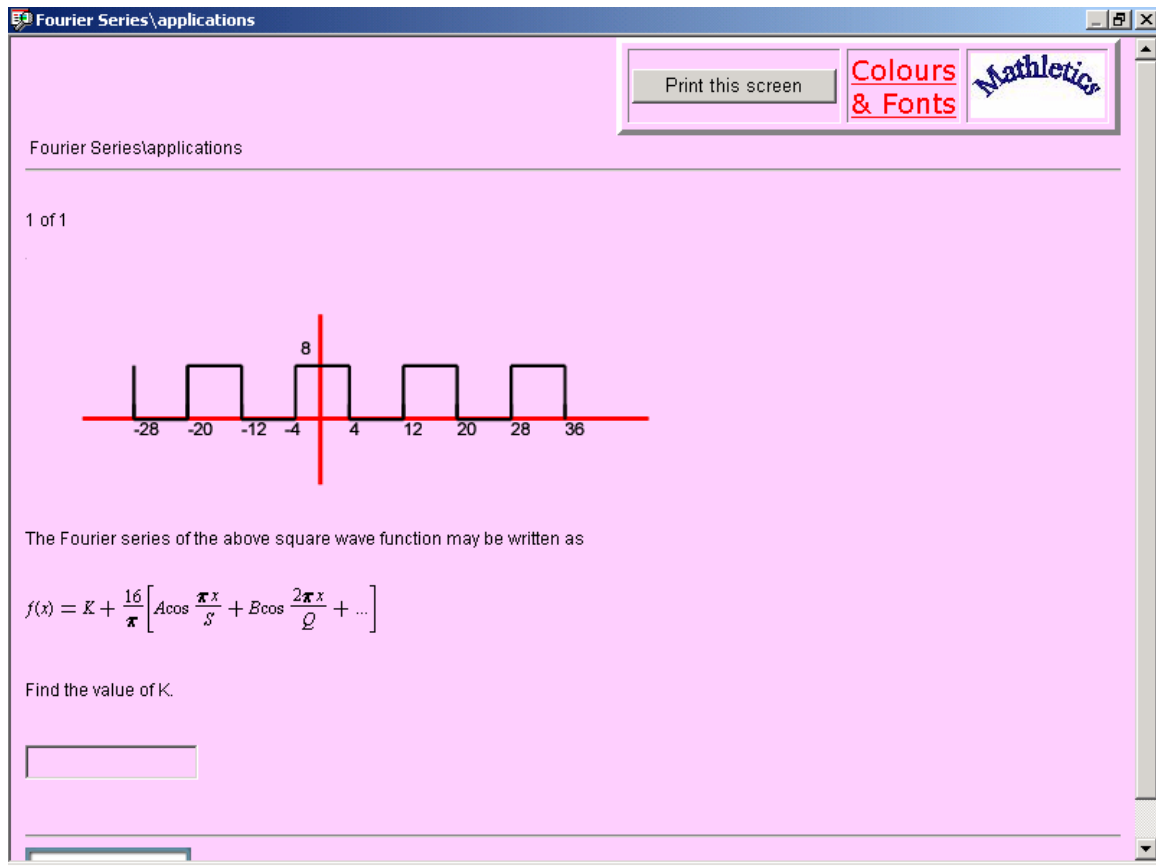
**Laplace transform\Properties - Name the correct property; 4YNSP**

Various properties of Laplace transform have been stated below.

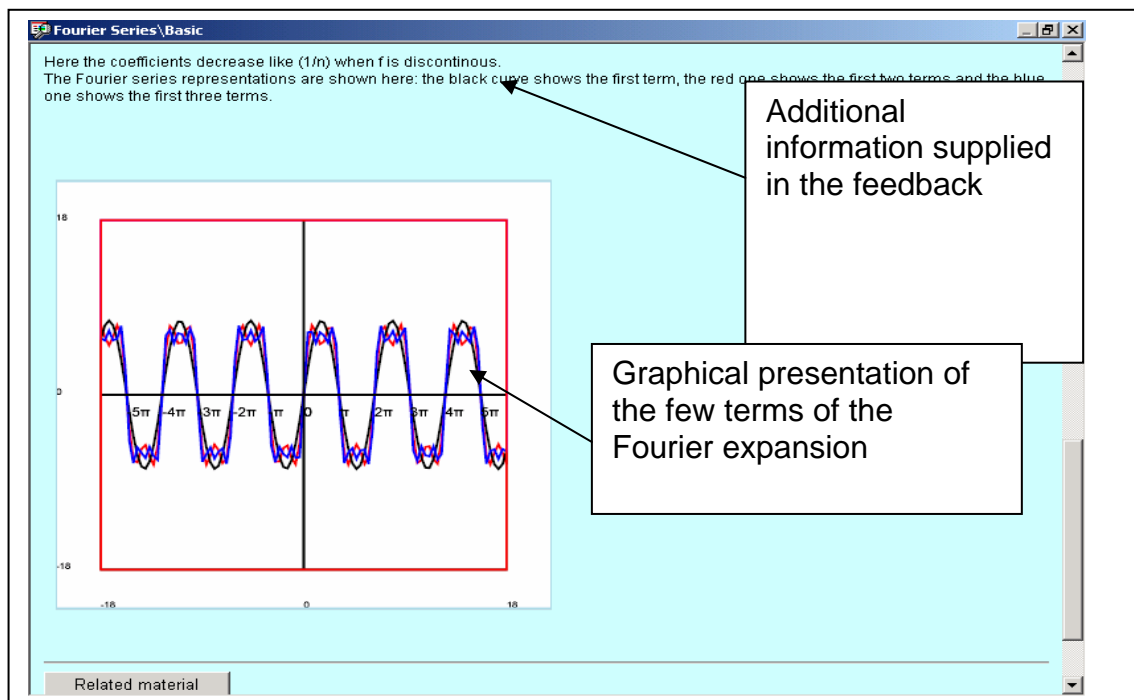
Has the right name of the property been used?  
 If you think so **input Y**.  
 If you think not **input N**

Properties	Y or N ?
According to integral theorem if $\mathcal{L}^{-1}[f(s)] = F(t)$  $\mathcal{L}^{-1}[f^n(s)] = (-1)^n t^n F(t)$	<input type="text"/>
According to differentiation theorem  and $\mathcal{L}^{-1}[g(s)] = G(t)$ then  $\mathcal{L}^{-1}[f(s)g(s)] = \int_0^t F(x)G(t-x)dx$	<input type="text"/>
According to convolution theorem $\int_s^\infty f(x)dx = -\frac{1}{t}F(t)$	<input type="text"/>
According to 2nd shifting theorem $\int_t^\infty f(x)dx = \frac{1}{t}F(t)$	<input type="text"/>

**Figure 1 Screen shot of a yes/no question for assessing general theorems and properties.**



**Figure 2 Screenshot of a question requiring analysis of a diagram.**



**Figure 3 Screenshot of part of the feedback using a graph. In contrast to the diagram in figure 2, which is really a schematic, this graph is accurately drawn.**

## **Trials and student feedback**

Questions spanning various learning levels for Laplace transforms and Fourier series were administered to second-year undergraduate students in three different tests in the months of October and November of 2006. The answer files have been analysed in an attempt to understand the questions' impact on the students' learning. Some new mal-rules were also identified through the analysis and were used in the construction of further questions on Fourier transforms.

All the questions have discrimination indices above 0.2, which indicates that no questions were invalid. The average facility values were around 0.5, indicating that the questions were of medium difficulty. However a certainty-based numerical input question and a true/false question had very low facility in comparison to the overall facility of the tests. The average facility value of the multiple-choice questions was more than that of the other type of questions. This probably reflects the effect of the information displayed on screen that allows students to check their answers against the options before clicking 'submit'. Whilst this casts doubt on using this question type for summative or mastery testing of students, for formative testing it is felt that, coupled with the very full feedback available, multi-choice questions are an effective way of building students' confidence.

Not surprisingly, students performed better in the lower-level questions than in the higher-level questions; those who were less certain scored lower than those who were more certain; and students did better in questions that tested a single concept than multi-concept questions. An exception to this appears to be that students were less able to identify general properties than apply them in specific examples. This may be due to such general properties being stated in a more abstract and mathematically terse way, or it may indicate deficiency in the conceptual learning of the topic. At the other end of the taxonomy, most mistakes occurred due to procedural errors, especially in the lower-level questions.

Results from a questionnaire suggest that students found the tests, and especially the feedback, useful. The marking scheme of some of the multiple-response questions has been set so that marks are obtained only if all the correct answers are chosen without choosing any incorrect options. About a quarter of the students considered this was not fair.

## **Conclusions**

Whilst questions at the lower levels of a modified Bloom's taxonomy can be created and shown to be effective in testing basic, albeit necessary, skills, any course in advanced calculus involving such topics as Fourier series or Laplace transforms will need the assessment of higher-, or *create*-level, questions. We give examples of how this can be done and the sort of feedback that should be offered to reinforce the learning of conceptually-difficult material. Generally multi-choice or numerical input type questions

(which serve well at lower levels) need to be augmented with other question types and/or question stem design that requires students to extract relevant material themselves, for example from a diagram. Trials have shown that whilst all of our questions were valid, some were perceived as unfair by students. Moreover, the success rate of different question types (as measured by question facility) was variable, with a new type of yes/no or true/false question testing general concepts or theorems proving to be challenging. Thus the choice of question type is important, especially in high-stakes assessments.

## References

- Ellis, E. 2006 Coding for an SVG graph plotter (private communication).
- Ellis, E., Greenhow, M., Hatt, J. 2006 Exportable technologies: MathML and SVG objects for CAA and web content *Proc 10<sup>th</sup> CAA Conf, Loughborough, July*. <http://www.caaconference.com/>
- Gardner-Medwin AR & Gahan M. 2003 Formative and Summative Confidence-Based Assessment, *Proc. 7<sup>th</sup> International Computer-Aided Assessment Conference, Loughborough, UK, July 2003*, pp. 147-155 <http://www.caaconference.com>
- Gill, M. & Greenhow, M. 2004, Setting objective tests in mathematics using QM Perception *Proc 8<sup>th</sup> CAA Conference, Loughborough, July* <http://www.caaconference.com>
- Greenhow, M. 1996 Computer based diagnostic tests and assessment at Brunel University, Published: Quarterly Newsletter of CTI Maths and Stats Vol 7 no 3 Aug 1996, pp 20-24.
- Hatt, J. & Baruah, N. 2006 The Reconceptualisation of The Revised Bloom's Taxonomy for Use in Mathematics and its Implementation into QM Perception and Mathletics, PRHE Conference, Liverpool, May. <http://hopelive.hope.ac.uk/PRHE/>
- Orton, A. 1983 Students' understanding of integration. *Educational Studies in Mathematics*, 14, 1-18
- Schechter, E. 1994 The Most Common Errors in Undergraduate Mathematics, <http://www.math.vanderbilt.edu/~schectex/commerrs/>
- Turns, J., Atman, C. J., and Adams, R. 2000 "Concept Maps for Engineering Education: A cognitively motivated tool supporting Varied Assessment Functions" *IEEE Transactions on Education*, v. 43, no 2, pp. 164-173.

**A STUDY INTO THE USE OF  
COMPUTER AIDED ASSESSMENT  
TO ENHANCE FORMATIVE  
ASSESSMENT DURING THE EARLY  
STAGES OF UNDERGRADUATE  
CHEMISTRY COURSES**

**Simon Bedford and Gareth Price**





# **A Study into the Use of Computer Aided Assessment to Enhance Formative Assessment during the Early Stages of Undergraduate Chemistry Courses**

Dr Simon B Bedford and Dr Gareth J. Price, Department of Chemistry, University of Bath, BATH, BA2 7AY, UK

[S.B.Bedford@bath.ac.uk](mailto:S.B.Bedford@bath.ac.uk)

01225 386143

## **Abstract**

A Virtual Learning Environment (WebCT and latter Moodle) was used to provide students with instant, meaningful feedback on their study of chemistry units during their first semester at University. Short multiple choice questions (MCQ's) were written covering each segment of material delivered in lectures and made available to students over the University computer intranet to allow "24/7" access. The most important aspect of the work was the feedback offered to students within the questions, which was written by undergraduate students to ensure its usefulness. The vast majority of the cohort used the MCQ's, most to gain formative feedback and some as a revision aid prior to summative examinations. During the evaluation, students reported that they found the ready access useful and helpful in learning the material. Some students used the MCQ's in preference to visiting tutors face to face (f2f) but most expressed a preference for the usual tutorial programme over such CAL methods. Most of the cohort used the feedback from the MCQ's to guide their revision, but again were not prepared to use CAL to replace f2f contact with tutors. Our work meets a number of the published conditions for effective feedback to occur. For example, it is immediate, timely and allows students to receive frequent feedback at a level which means that it can be used to inform further study. In the first year of using the MCQ's, there was a significant increase in the average marks in the end of unit examinations and a decrease in the drop-out rate during Semester 1. Although firm conclusions cannot be drawn from one year's data, these results together with the very positive reaction from the students encourage us to further develop the approach into the open source VLE Moodle, which allowed us to address some of the issues.

## **Introduction and Rationale**

A number of staff in the department were concerned that UG students were not fully engaging with the programme of workshops and tutorials and so were not receiving useful formative feedback until end-of-semester examinations.

By this time it was often too late to fill gaps in knowledge or to correct misunderstandings since the teaching programme (which builds on this work) moves on at an increased pace. We were anxious to overcome this while not “spoon-feeding” students; we needed a method that would enhance and encourage them to take responsibility for their own learning and adopt a student centred approach. Although we are new to CAA in general, a small number of colleagues were keen to get involved. We had some experience in using a computer based question program (Question Mark Perception) but, for other purposes, were trialling a VLE and so were keen to investigate whether this could help us. All first year students live in University accommodation that is networked so allowing ready access to CAL materials. The University has a Learning Centre with > 450 networked PC's which is open 24 hr per day. It therefore seemed to us that CAA would potentially allow ready access to feedback.

*In terms of the conditions for successful feedback, those most directly relevant to this project were:*

1. Sufficient feedback is provided, often enough and in the appropriate detail
2. The feedback is provided rapidly to be useful to the learner
3. Feedback focuses on learning rather than on ‘marks’.
4. Feedback is understandable to students, given their sophistication
5. Students should act upon the feedback in order to improve their learning.

The vast majority of students in this study were school leavers with A-level grades in the range BCC – AAA. Around half-a-dozen held International Baccalaureate qualifications, two progressed from university Foundation courses designed for broad entry to HE and one from a GNVQ route. In this cohort, there were no students older than 25. Approx. 40% of the cohort was female. Chemistry teaching at Bath is based around a traditional lecture format (ca. 6 per week, 50minutes duration) supplemented by problem classes (2 per week, 50minutes duration) and small-group tutorials (1 per week, 50minutes duration) with 5 – 6 students in each. Most formative feedback was obtained by students during tutorial and workshop sessions.

## **Methods**

The project background was largely developed through informal discussions with students during tutorials and with colleagues. More in depth discussions were held with a small number of students who had recently completed their first year to further refine our ideas. However, at this stage “data” were largely anecdotal. For each small section (2 – 5 hrs) of lecture material, a short series of multiple choice questions were written to allow students to test their basic understanding of the fundamentals of the material as well as to give some questions to determine whether they could apply this knowledge. This was mounted on the University computer network and students encouraged to use

it during their studies in order to monitor their progress. It was in no way compulsory for students. However, part of the summative assessment for the units is a 2 hr MCQ unseen examination and students were told that most of the “past paper questions” were included in the MCQ’s. Individual MCQ’s were ‘released’ as the material was covered in lectures during Semester 1. A range of different question types was employed to test knowledge, ability to interpret simple observations as well as background mathematical skills and quantitative abilities. One advantage of using a computer over a paper based system is that some questions were designed around animations to enhance students understanding of e.g. reaction mechanisms. (Examples of the questions and the approaches are available on request). Simply telling students whether they had answered questions correctly or not would be of limited value. Into each question was therefore built some constructive feedback. Even if the question was right, feedback was given to enhance the learning (e.g. “Well done – you obviously remembered the correct units for the gas constant, R”) and reinforce good habits. Wrong answers were met with an attempt to indicate where students had made errors.(e.g. “Have you considered the units of the gas constant ?”, “Think about how many joules are in a kilojoule” or “What does the ‘1’ in ‘SN<sub>1</sub>’ mean?”. In this way, students were not simply fed the answer but forced to think about why they were not correct in the first attempt. In the event that they were completely unable to answer a question, students were encouraged to use the question as a basis for discussions during tutorials and workshop sessions. The ready access to the computer network facilitated several conditions. No marks were recorded by staff (although they are available within the VLE) so that students were aware that doing the MCQ was solely to check their current state of knowledge and ability and for them to gauge areas of weakness on which further work was needed. In order to meet Condition 4, a student was employed who had just completed the year of study. They wrote or edited much of the feedback to ensure that it was at the correct level.

### *Resources*

We used a VLE – WebCT, and later on integrated it into Moodle to make use of resources such as wiki’s and synchronous discussion forums. In principle any CAA system (e.g. Question Mark Perception, etc.) could be used but we were evaluating a VLE for other uses and it was convenient for students to only use one system. Students need to be able to use a PC in order to access the VLE. A crude evaluation of the effectiveness of our approach can be gained from a comparison of the 2004/05 unit results and the number of students who dropped out during Semester 1 compared with previous years. However, this of course is open to very considerable uncertainty given the number of factors that influence these criteria. The primary evaluation has therefore been by asking students to fill in a questionnaire (see appendix for paper based version). In addition, our project was aimed at students right at the start of their university careers so that they would not have had time to develop study strategies sufficiently early to make a later comparison meaningful. Also, we wanted only to use one questionnaire so as to avoid “questionnaire fatigue”. A feedback questionnaire was therefore designed to incorporate the relevant questions directly relating to our project and more

generic ones about the VLE. These were produced in both paper and e-reports media. Students were asked to complete the questionnaire in early April, allowing time after the examinations and receipt of results (mid February) for students to reflect on their use of our MCQ's. The results of the questionnaire are shown in Appendix 2. Out of a total cohort of 115, 98 students returned questionnaires, a response rate of 85%.

## Results and Discussion

In terms of the summative assessment of the unit, there was a distinct improvement in performance for this session. The assessment comprises a piece of coursework done mid-way through the semester together with a MCQ examination and a problems based examination held at the end of the semester. This year's cohort showed a significant improvement over the previous year with the average mark moving from 56.7 (s.d. = 13.4 to 65.2 (s.d.=10.6) this year. For each individual component, an improvement was shown with the most pronounced (perhaps not unexpectedly) in the MCQ examination where the average moved from 53.1 to 60.1. In the current academic year, only 1 student withdrew from the course before the Easter vacation compared with 6 in the previous session. Of course this is at best a crude evaluation of the effectiveness of our approach. Many other factors affect performance and withdrawal rates. The average A-level entry grades were somewhat higher for the later cohort (BBB versus BBC) and this may account for some of the improvement. However, we can at least conclude that the introduction of enhanced feedback has not had a negative effect on performance

### *Analysis of evaluation questionnaires*

Of the cohort who answered the questionnaires, we were pleased to see that over 80% had used to the system to at least some extent. Given the well known cynicism of some students (the "it doesn't count so I won't bother" syndrome) this was satisfying. Of the students who did not use the packages, (18% of the respondents), their quoted reasons can be grouped into three main categories:

#### *1. Motivation and student effort, typified by responses such as:*

- "Didn't have the time, kept forgetting."
- "Didn't think it would be worthwhile".
- "General laziness. I also found them a little tricky to find. Lots of good intentions but never got around to it!"
- "Never had time during the exam period, spent most time on past papers etc. Should have planned to use them earlier in the term."

We have to accept that some students will never take advantage of the learning opportunities offered no matter what the mode of delivery.

## *2. Technical factors, including:*

- “tried it a couple of times didn’t work, so couldn’t actually use it. Kept freezing. If it had worked would have used it.”
- “Couldn’t find them on the net. More links from the Chemistry pages would be helpful.”
- “also would have had to have gone to the library in order to use a computer.”
- “I did not have computer access in my room and it can be difficult to get a computer in the library.”
- “Attempted to use them but became frustrated with systems’ inability to handle 99% correct answers. e.g. 99kJmol<sup>-1</sup> was right but 99(space)kJmol<sup>-1</sup> was wrong.”

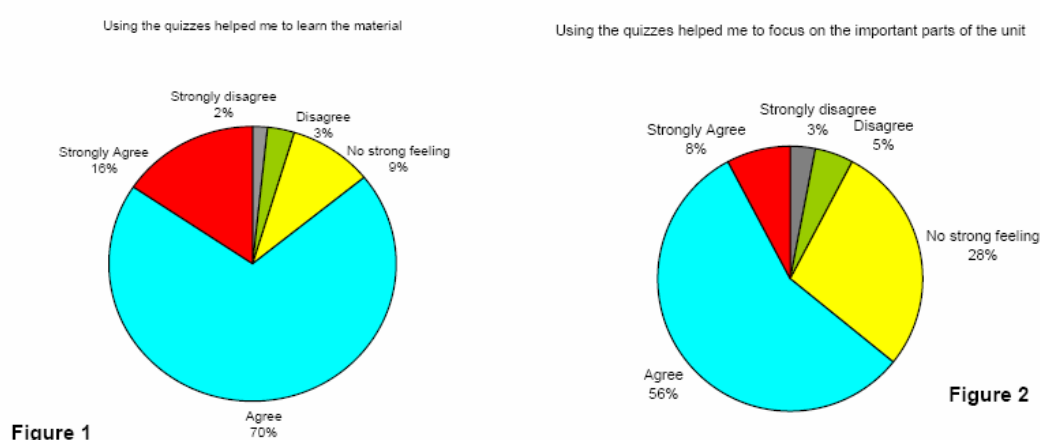
This was a relatively small number of reported problems considering it was our first experience of using the VLE system. The access problems are something that we will take seriously, and were generally down to linking our VLE with our student records system (SAMIS). The final comment is interesting but development of CAA systems has now rectified this. However, it seems that this student was focussing more on ‘getting the mark’ than acknowledging that they had obtained the right answer as an aid to learning.

## *3. Pedagogic factors and preferred learning and revision styles:*

- “I don’t find computer learning particularly useful. I tend to remember things by rote if I use MCQ’s, instead of learning and understanding. Part of this was due to lack of time – I prioritised that my normal revision method was more effective.”
- “Preferred to revise using books and notes with past papers, rather than using the computer, I don’t really feel that MCQ’s are my favourite way to learn, I often feel extremely unmotivated to do them.”
- “I did not feel that the MCQ’s would help me, as they are not the style of revision that I know helps me the most.”
- “I would rather learn using a pen and paper! “
- “I find past exam papers more useful because in the past, MCQ’s have not been as hard etc. as past papers.”
- “I used past exam questions, as well as tutorials and workshops to assess how well I revised.
- Also, I didn’t judge quite how much revision was needed in order to do well, and was fairly lazy!”

Given the strong steer from many sources that current students are computer literate and regard traditional teaching such as “chalk and talk” as old-fashioned, we were surprised at these comments, albeit that they are a small number. The responses were initially anonymous so that it is not possible to correlate use of the system with individual comments to see if students’

performance might have been hampered by not using the MCQ's. Although the evidence from the latest study using Moodle seems to show a strong link. Of the 80 students who did use the system, 65% used them for formative feedback during the semester, the other 35% using them as a revision tool in the run up to the end of semester examinations. Of the former group, about half used all the MCQ's and of the rest, the preference was to use the MCQ's for units that were found difficult rather than those in which students were most interested (questions 2 and 3). Few students used them only to prepare for coursework. A gratifying feature was that the majority of students felt that using the MCQ's had helped them to learn the material covered in the units (see Figures 1 and 2). While anecdotal in nature this, along with the improvement in examination performance, suggests that we met condition 5.



Significantly though, students were neutral on whether the feedback had helped them plan their study (question 13). Only 7 students either strongly agreed or strongly disagreed that this was the case and equal numbers either agreed or disagreed (Figure 3). Similar responses were received concerning the effectiveness of the approach in bridging the school-university transition (question 14). There was a slight preference for the suggestion that using the packages helped to develop independent learning although few students seemed to have used the feedback as a basis for seeking further help during tutorials. Only 10 students felt that the CAA approach was better than the traditional tutorials, even though it is more readily available (Figure 4).

I used the feedback to help me plan my study of the units

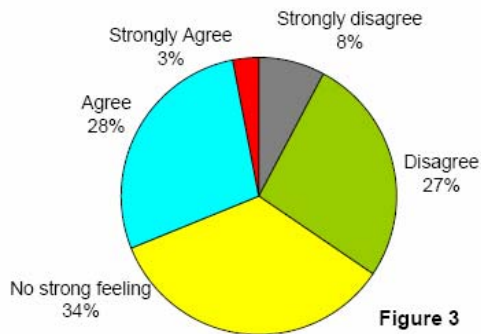


Figure 3

This type of instant feedback is better than having tutorials

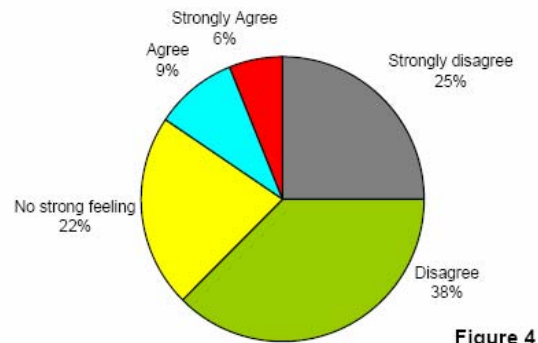


Figure 4

Only 5 students felt that the questions were too hard and 11 did not understand the feedback given. While the latter figure is higher than we would like, the results indicate that we largely met our target of the MCQ's and feedback being at the correct level for the particular cohort of students, meeting condition for effective feedback.

A larger proportion of the class used the feedback MCQ's as an aid to revision for the final assessments. Of these 80 students, all but 9 used the MCQ's to gauge how their revision was proceeding and the majority used them as a diagnostic tool to focus their revision (Figure 5) and the majority (73%) agreed that the feedback was helpful in learning the material. 85% of students liked the ability to get answers at any time, of relevance to Condition 2. Again not surprisingly, students expressed strong preference for visiting Tutors to get problems answered rather than simply using electronic means (Figure 6).

I used the feedback to tell me where to spend most of my revision time

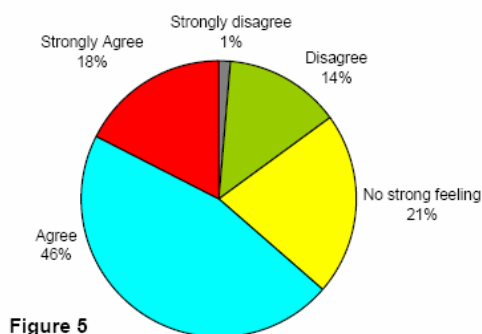


Figure 5

I would prefer to visit my tutor/lecturer to get questions answered

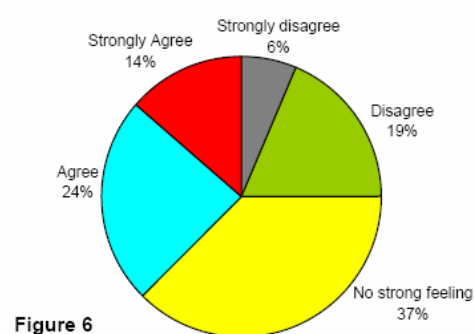


Figure 6

## Analysis of overall aims and objectives of the project

When we designed the system, our hope was that the system would lead to:

- All students using it after each lecture “section” was completed
- Better focus of tutorials and workshops
- Prevention of some visits to staff with trivial problems
- More effective use of staff time in dealing with problems
- Less questions to staff during the revision period
- More effective revision
- Better performance in assessments

So, what was the result? A good proportion (82% of a 85% response rate) of the cohort did use it for formative feedback during the semester while a second group used it as a revision aid. Although few students said that they used the MCQ’s to focus tutorials, comments from staff suggested that there were less visits with trivial problems this year although there is no firm evidence. Most students felt that using the MCQ’s had improved their overall assessment performance and this is supported by the change in average marks, albeit for a single cohort.

Our aim was to use CAA to enhance our traditional teaching methods, not to replace them. In this we seem to have been successful, at least in terms of student acceptability. One telling comment which applies to CAL methods in general rather than specifically to this project was:

*“I came to Bath because of the friendliness and approachability of staff – and then you send me away to work with a computer on my own”.*

Clearly, we need to manage the introduction of CAL carefully if detrimental changes to our departmental ethos are not to occur.

### Analysis of the conditions for effective feedback

The conditions of major interest to this project are shown in **bold**.

	Condition	Project response
<b>1</b>	Assessed tasks capture sufficient student time and effort	The MCQ’s were well used and so captured time and effort. The results suggest that this was, in the main, sufficient.
<b>2</b>	These tasks distribute student effort evenly across topics & weeks	This applies to those students who used the feedback through the semester, less so for those using it as a revision aid.
<b>3</b>	These tasks engage students in productive learning activity	The results suggest that activity to have been productive!
<b>4</b>	Assessment communicates clear and high expectations to students	Not applicable here.
<b>5</b>	<b>Sufficient feedback is provided, often enough &amp; in enough detail</b>	<b>Feedback is available whenever students want it; it is up to them to use the MCQ’s. Most students found the level of detail in the feedback appropriate.</b>



<b>6</b>	<b>The feedback is provided quickly enough to be useful to students</b>	<b>It is instant and so can be acted upon rapidly.</b>
<b>7</b>	<b>Feedback focuses on learning rather than on marks or students</b>	<b>The feedback focuses on getting students to think about the material and to re-study in the case of incorrect answers. There are no links to assessment grades.</b>
<b>8</b>	Feedback is linked to the purpose of the assignment and to criteria	Not applicable here.
<b>9</b>	<b>Feedback is understandable to students, given their sophistication</b>	<b>The feedback was designed by students and the survey results suggest that it was at the right level.</b>
<b>10</b>	Feedback is received by students and attended to	Feedback is certainly received by students and their comments suggest that most acted on it.
<b>11</b>	Feedback is acted upon by students to improve their work or their learning	This is difficult to quantify but seems to have been a satisfactory result of our work.

## Conclusions

Overall, the project was successful. We underestimated the time commitment required to set up such a system of MCQ's, even when using a commercial software product such as WebCT or open source Moodle and importing questions into it from WebCT. We were pleased at the comparative lack of technical problems faced by students – albeit that this was offset by the staff set-up time spent ensuring that things were robust. The main unforeseen circumstance that we encountered was the comparative overloading of students in the first few weeks of their university careers. Although we hoped that our feedback system would help in the school-university transition, it was hardly used in the first few weeks. Enquiries to students showed that many were overwhelmed by the number of new procedures, tasks, skills and general activities that take place in the first couple of weeks, both academically and socially. A second introductory session was held after 4-5 weeks of the semester and usage increased afterward. The initial set-up time and technical support necessary for such a system should not be underestimated. Sourcing, devising and inputting the questions was time consuming (ca. 13 weeks for an undergraduate student). Even though a commercial VLE was used initially, there were technical issues in its use in terms of student access, passwords etc. and in working out how to include some question types (e.g. those with video clips or the interface with PowerPoint). Individual students also needed help with accessing and navigating the system, although this improved when Moodle was adopted with its user friendly interface.

## **References**

Conditions Under Which Assessment Supports Students' Learning Graham Gibbs and Claire Simpson Learning and Teaching in Higher Education 1, 3-31.

## Appendix 1: Student evaluation questionnaire

### FAST Project Feedback Questionnaire

The feedback packages were developed using funding from a UK wide initiative entitled *Formative Assessment in Science and Engineering*. Your feedback will be important in refining the quizzes for future students and will feed into the national evaluation of the project.

#### Section 1: Your use of the quizzes

Did you use the WebCT revision/feedback quizzes?

☐ Yes (Please go to Section 2)

☐ No (Please go Section 4)

#### Section 2: Using the quizzes for feedback

(If you used the quizzes only in the run up to the end-of-unit examinations, please go to Section 3 overleaf)

*Please rate the extent to which you agree with each of the following statements by ticking the appropriate box*

	strongly disagree	disagree	neither agree nor disagree	agree	strongly agree
I used <u>all</u> the quizzes that were provided	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I used the quizzes only for units that I found difficult	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I used the quizzes only for units that interested me most	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I took the quizzes several times to see if I improved	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I used the quizzes only to help prepare for coursework/tutorials	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Using the quizzes helped me to focus on the important parts of the units	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Using the quizzes helped me to learn the material	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
The questions were too hard	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I couldn't understand the feedback that was given in the questions	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
This type of "instant feedback" is better than having tutorials	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
The feedback was sufficient to help me understand where I went wrong	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I used the feedback to help me to plan my study of the units	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Using the quizzes helped me to bridge the gap between school/college and university	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Using these packages helped me to develop independent learning rather than just relying on tutors

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
--------------------------	--------------------------	--------------------------	--------------------------	--------------------------

I used the feedback to know what questions to ask during tutorials

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
--------------------------	--------------------------	--------------------------	--------------------------	--------------------------

*Please add any comments that you wish to make on the reverse of the questionnaire.*

### Section 3. Using the quizzes for exam revision

*Please rate the extent to which you agree with each of the following statements by ticking the appropriate box*

Strongly disagree  
disagree  
No strong feeling  
agree  
Strongly agree

I used the quizzes to assess how well my revision was going

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
--------------------------	--------------------------	--------------------------	--------------------------	--------------------------

I used the feedback to tell me where I needed to spend most of my revision time

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
--------------------------	--------------------------	--------------------------	--------------------------	--------------------------

The feedback on answers was useful in learning the material

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
--------------------------	--------------------------	--------------------------	--------------------------	--------------------------

Using the quizzes meant that I didn't have to visit my tutor/lecturer with problems

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
--------------------------	--------------------------	--------------------------	--------------------------	--------------------------

I liked being able to get answers at any time

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
--------------------------	--------------------------	--------------------------	--------------------------	--------------------------

I would prefer to visit my tutor /lecturer to get my questions answered.

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
--------------------------	--------------------------	--------------------------	--------------------------	--------------------------

Having access to the quizzes improved my performance in the examinations.

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
--------------------------	--------------------------	--------------------------	--------------------------	--------------------------

*Please add any comments that you wish to make on the reverse of the questionnaire.*

### Section 4.

Please explain why you didn't use the quizzes.

*(Continue overleaf if necessary)*

You have finished the questionnaire. *Thank you for your help!*

## **Appendix 2: Evaluation questionnaire responses**

Section 1		NO	(%)	YES	(%)	Strongly Agree	Agree	No strong feeling	Disagree	Strongly disagree	TOTAL
1 Did you use the quizzes		18	18.4	80	81.6						98
Section 2											
2 I used all the quizzes that were provided		12	20	9	13	10				64	19%
3 I used the quizzes only for units that I found difficult		7	17	5	29	6				64	11%
4 I used the quizzes only for units that interested me most		11	30	12	10	1				64	17%
5 I took the quizzes several times to see if I improved		7	21	11	21	4				64	11%
6 I used the quizzes only to help prepare for coursework/tutorials		13	30	14	6	1				64	20%
7 Using the quizzes helped me to focus on the important parts of the unit		2	3	18	36	5				64	3%
8 Using the quizzes helped me to learn the material		1	2	6	45	10				64	2%
9 The questions were too hard		8	28	23	5					64	13%
10 I couldn't understand the feedback that was given in the questions		8	37	8	10	1				64	13%
11 This type of instant feedback is better than having tutorials		16	24	14	6	4				64	23%
12 Feedback was sufficient to help me understand where I went wrong		3	9	17	32	3				64	5%
13 I used the feedback to help me plan my study of the units		5	17	22	18	2				64	8%
14 Using the quizzes helped me to bridge the gap between school/college and university		8	16	23	17					64	13%
15 Using these packages helped me to develop independent learning		3	9	27	22	3				64	5%
16 I used the feedback to know what questions to ask during tutorials		8	19	23	14					64	13%
Section 3											
17 I used the quizzes to assess how well my revision was going			9	10	44	17				80	0%
18 I used the feedback to tell me where to spend most of my revision time		1	11	17	37	14				80	1%
19 The feedback on answers was helpful in learning the material		1	4	17	46	12				80	1%
20 Using quizzes meant I didn't have to visit my tutor/lecturer with problems		12	32	18	17	1				80	15%
21 I liked being able to get answers at any time		1	1	10	39	29				80	1%
22 I would prefer to visit my tutor/lecturer to get questions answered		5	15	30	19	11				80	6%
23 Having access to the quizzes improved my performance in the exams		1	7	33	31	8				80	1%



# **IMPLEMENTING CAA IN CHEMISTRY: A CASE STUDY**

**Emilia Bertolo, Glenis Lambert**





# Implementing CAA in Chemistry: A Case Study

Emilia Bertolo,<sup>a</sup> Glenis Lambert<sup>b</sup>

<sup>a</sup> Department of Geographical and Life Sciences, Canterbury  
Christ Church University, Canterbury, Kent, CT1 1QU (UK)  
[meb27@canterbury.ac.uk](mailto:meb27@canterbury.ac.uk)

<sup>b</sup> Learning and Teaching Enhancement Unit, Canterbury Christ  
Church University, Canterbury, Kent, CT1 1QU (UK)

## Abstract

Computer aided assessment (CAA) was implemented in the level 1 module Skills for Forensics Investigators; the assignment was focused on several chemistry concepts. The aim was to provide students with rapid feedback, while trying to enhance their engagement with the subject; reducing the lecturer's marking load was perceived as an added bonus. The CAA system used was Perception from Question Mark Computing; the assessment comprised two components, one formative and one summative. The formative test could be accessed at any time, and provided feedback that sought to guide further learning; the summative component had no feedback and could only be taken once. From the lecturer's perspective, the experience was very positive. The initial time invested preparing the assessment was considerable; however, that time was used in a creative way (designing the assignment) as opposed to a conventional paper based assessment, in which the time would be spent in routine marking. A total of 83 students, 94% of the students for that module, participated in the assessment process, suggesting that the use of technology did not prevent students from taking the assignment. Student evaluation was gathered via anonymous on-line questionnaires; 38.5 % of all the students involved in the assessment (32 students) answered the evaluation survey. Results indicate that the CAA system has made a positive impact upon the students' learning experience. This assessment raised some issues regarding students' "last minute" working practices. Students who left the test until the last minute and who experienced difficulties were dealt with individually, but this is an aspect which needs to be resolved through clear regulations rather than on an ad-hoc basis. Overall, the experience has proven very positive for both staff and students. The success of this assignment has led to improved communication with the students on the nature of their online assessment.

## Introduction

Computer assisted or computer aided assessment (CAA) refers to the use of computers in assessment (Gladwin, 2005). Jenkins (2004) has compiled a range of case studies illustrating the potential benefits and limitations of CAA. Among the benefits, Jenkins identifies repeatability, close connection between the activity and the feedback, flexibility of access and increased student motivation. As pitfalls, the author mentions development time, potential risks associated with hardware, software and administration, and the necessity for students to possess appropriate computing skills and experience. Wisely used, CAA can be far richer than paper-based assessment and have a very positive influence in the assessment process, offering quick, often instant, marking and feedback (Bull and Danson, 2004). Flexibility is an added bonus for the case of open access web-based assessments, since the tests can be taken at a location and time to suit the student (Bull and Danson, 2004).

Considerable efforts have been made to introduce CAA in chemistry at HE level, with the HE Academy playing a pivotal role in many cases. Adams et al (2002) have developed a series of online question banks using the QuestionMark Perception assessment management system. Price (2006) conducted a project aimed to enhance students' early experience at university, using CAA to provide formative feedback to students in their first year of undergraduate Chemistry courses. Over 80% of the cohort used the quizzes, and students reported that they found the ready access useful and helpful. Lowry (2005) used a CAA system for formative self-assessment, to provide chemistry support to Environmental Science students. Most students traditionally consider chemistry as "hard" science, and have difficulties engaging with the subject; the premise was that any mechanism that increased students' interaction with chemistry would be beneficial. His study concludes that the CAA system made a positive impact upon the learning experience of the students involved.

### *Rationale behind this case study*

This paper describes the results of a computer assessment set up for level 1 students of the module Skills for Forensics Investigators (2005-2006 cohort). The assessment covered materials taught for a total of eight contact hours. The aim of the project was to use CAA to design an assessment procedure that could:

- provide a closer connection between the assignment and the subsequent feedback, and
- facilitate students' engagement with the subject.

The previous cohort had completed a paper-based assignment, consisting of short questions. Due to the large number of students involved (ca. 100), it was difficult to provide students with quick feedback. It was thus felt that the formative component, which should be part of any assignment, had not been sufficiently fulfilled. Moreover, informal feedback from the students had highlighted the difficulties for some students to engage with the subject, which can appear difficult and unattractive in some cases; this is a key problem

identified by other authors when teaching science subjects in an HE context (Overton, 2003; Lowry, 2005).

One of the advantages of CAA is that it can efficiently shorten the time gap between assignment and feedback. Moreover, Jenkins (2004) and Lowry (2005) both mention increased student motivation as a benefit of CAA. It was thus hoped that replacing the paper based assessment with a computer marked one would be positive for the students, with the added benefit of considerably reducing the lecturer's marking load.

From the lecturer's point of view, the key concerns were the initial time investment necessary to design the assessment, and the difficulty of designing pedagogically sound questions (see Clarke, 2001, and King et al, 2001). Since the learning outcomes that the assessment had to test were reasonably low level according to Bloom's taxonomy (identify, recall, calculate, etc.), it was not thought that the potential limitations of CAA in testing higher level learning outcomes to be a serious concern in this case. With the emphasis on facilitating students' engagement with the subject, and in view of the fact that the existing paper-based assessment was open, it was decided to go ahead, but maintain a careful watch on what was happening within the system to try to isolate any obvious malpractice.

## **Method**

### *The assessment procedure*

The CAA software used was Perception from Question Mark Computing. A total of 83 students (94% of the students registered on the course) participated in the assessment exercise, which ran over a two-week period. Access was not restricted to just the computers on campus and thus students could do the tests from home. The assessment consisted of two tests, one formative and one summative. The formative test could be taken several times; once started, it had to be completed within 30 minutes. The formative test was aimed to:

- avoid/minimise "computer anxiety": the format of the questions was similar to those of the summative test, so students could familiarise themselves with the various styles of questions;
- spot any unforeseen technical problems at the earliest opportunity;
- allow students to practice the concepts learned in the lessons, and enhance their knowledge of the subject: besides the mark achieved, students could access feedback relating to their answers; the feedback was constructed so that it tried to explain why an answer was incorrect, but not so that it gave the correct answer. The aim was to get the students to consider their understanding and not just to memorise the right answers.

The summative test could be only be accessed once, and did not provide feedback or the mark; once started, students had one hour to complete it. The question types used were multiple choice, fill in the blanks, true/false and

numerical, and some of the questions included images. The marks were released to the students just after the two-week assessment period had ended.

### *Student evaluation*

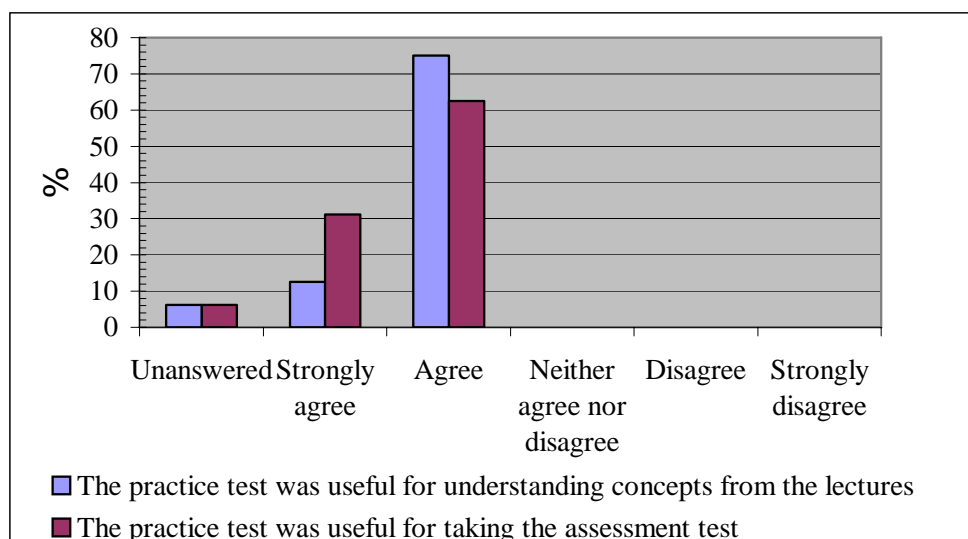
After the assessment had concluded, student evaluation was gathered via an on-line questionnaire. The questionnaire, modelled on the one used by Bullock (2001), comprised 10 questions, 9 based on a 5-point Likert scale and a final one to gather any further comment. From the 83 students that took part in the test, 32 responses were received (38.5%). The graphs constructed with answers to the 5-point Likert scale questions are shown in Figures 1 to 6; the further comments can be found in the results and discussion section.

## **Results and discussion**

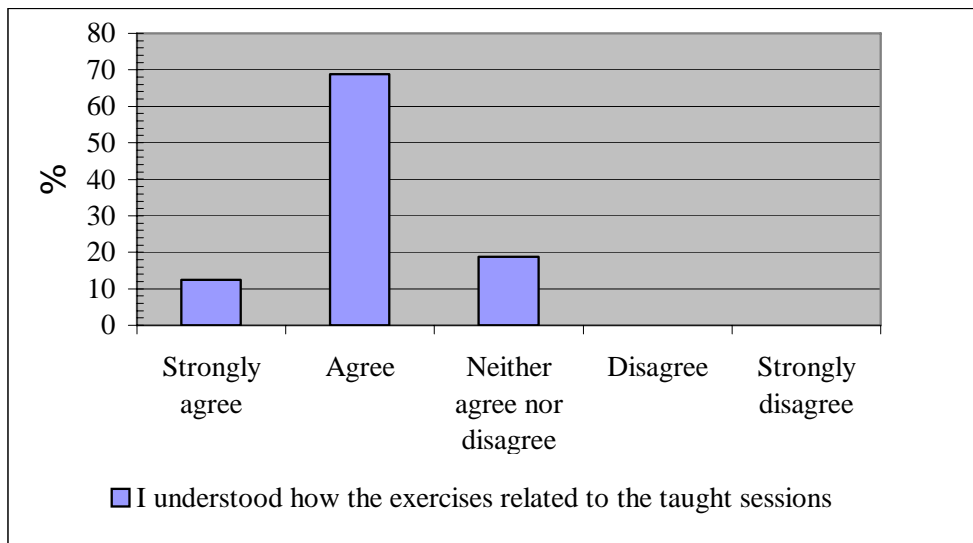
### *Student evaluation*

Some interesting observations can be made regarding students' perceptions about the experience, although the relatively low number of responses (38.5 % of all the students that took part in the assessment) is not enough to extrapolate conclusions to the whole group. The first two questions related to the usefulness of the formative test: Figure 1 shows that 93.7% of the students agreed or strongly agreed that the practice test was useful to prepare for the assessed test, and most students (87.5%) agreed that the practice test helped them understand the concepts explained in the lecture. Students also seemed to identify the link between the assessment and the taught sessions (see Figure 2).

**Figure 1. Student responses regarding the practice test, given as % (n = 32).**

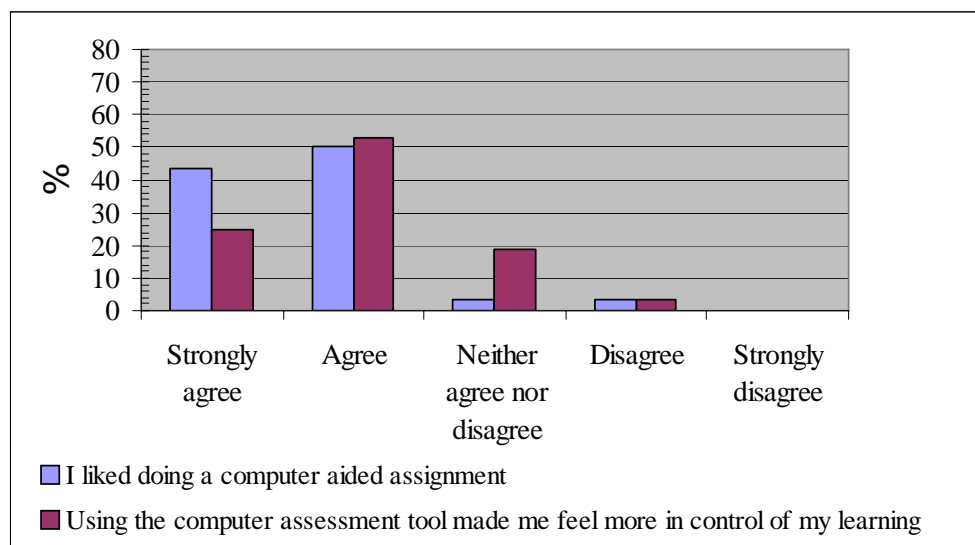


**Figure 2. Student responses, as a %, regarding the link between the assessment and the taught sessions (n = 32).**

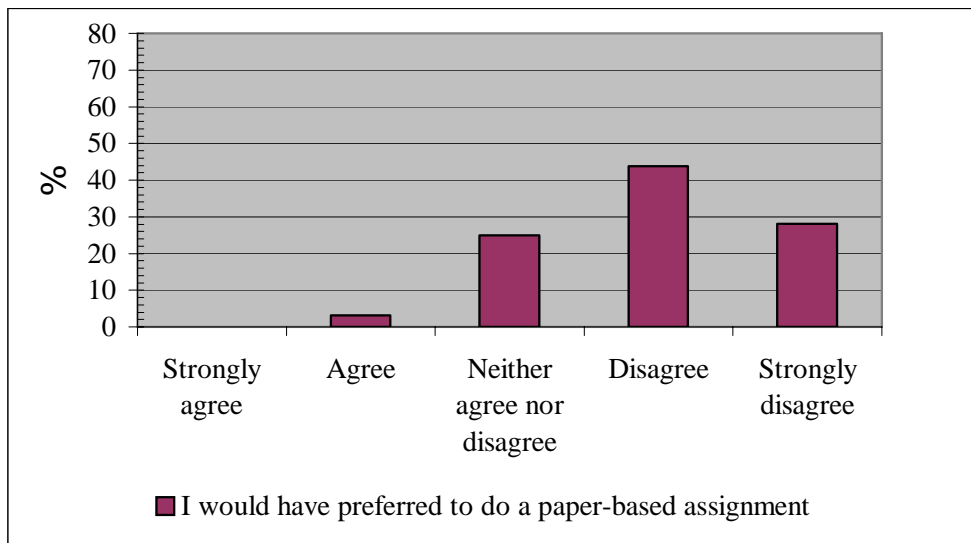


Figures 3 and 4 show students' responses regarding CAA vs conventional paper based assignments. In general, students' attitude towards the computer assessment was very positive. 93.7% of the students liked doing the CAA, and 78% of the students agreed/strongly agreed that using a computer tool made them feel more in control of their learning. Only one student would have preferred to do a paper-based assignment; 23 students (71.8%) disagree/strongly disagree with that statement, and the other 8 (25%) were neutral to that statement. Figure 5 shows students' views about the level of support received: only one student did not feel sufficiently supported, while the rest (46.8%) felt neutral about that point or agreed/strongly agreed (37.5%) they had been well supported.

**Figure 3. Student views on CAA, given as a % (n = 32).**



**Figure 4. Student preferences, as a %, regarding paper-based assessments for this module (n = 32).**



**Figure 5. Student views on the adequacy of the staff support received, given as a % (n = 32).**

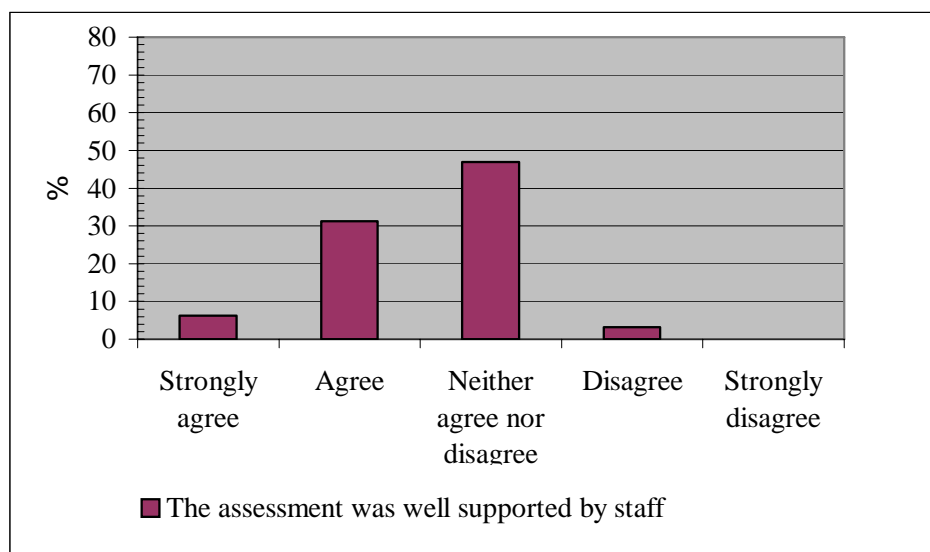
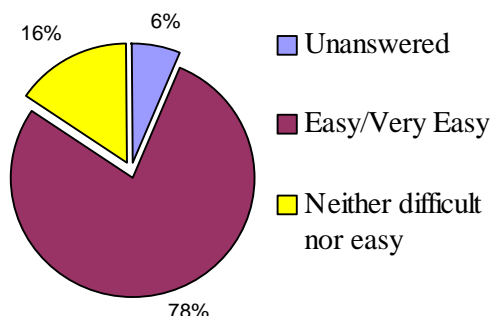


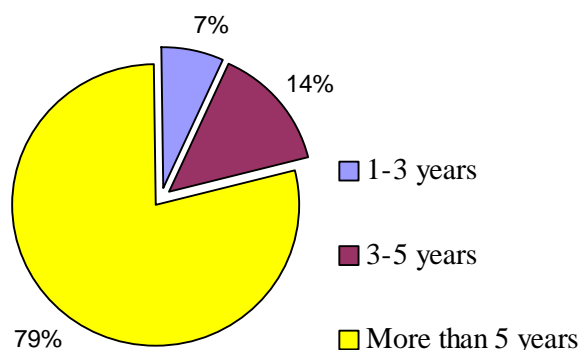
Figure 6 shows students' responses on the software used for the assessment and their previous computer experience. This group was mainly composed of experienced users, with 79% of them saying they had been using computers for more than 5 years. They did not seem to have problems with the software used in the assessment, with 78% saying it was easy/very easy to use. It is also noteworthy that 83 out of 88 students participated in the assessment process (94%), which suggests that the use of technology did not prevent students from taking the assignment.

**Figure 6. Student responses on a) the software used for the assessment and b) their previous computer experience**

a) How easy was the software to use?



b) How long have you been using computers?



Of the 32 responses received, only five students added a further comment. One was related to the lectures content (“Slight confusion on one of the questions about chromatography, maybe some sort of tutorial or helpful guide within the practice assessment could be used? Apart from that I found the assessment useful”), and the other four were focused on the assessment itself. In all cases, they were very positive about the experience (see below).

“I think that the online test programme is a very good idea, you can be tested and not be as nervous as you would sitting in a hall full of people”

“I preferred doing the computer aided test as it allowed me to do it in my own time”

“the test was useful and the practice test was extremely useful to help understand what the real test was going to be like”

“I have no further comments. I thought the on-line chemistry test was extremely useful and it helped me understand chemistry more”.

#### *The lecturer’s perspective*

Translating the original paper assignment into a computer one was not too difficult, since the original test consisted on a series of short answer questions and mathematical problems, but it was time consuming. Once the assessment was deployed, the process ran surprisingly smoothly considering the number of students involved. Only a handful of students reported technical problems. One issue not anticipated was the totally unrealistic expectations about staff availability held by some students: they seemed to think that, since the tests were available 24hr per day, the same would be true of staff. This could be why student perception of staff support is not as positive as it could have been expected. A possible way of addressing the problem would be to include details about staff availability in the assessment instructions.

### *The learning technologist's perspective*

In 2004, Canterbury Christ Church University began to use Questionmark/Perception for small scale implementation of medium-stakes summative assessments. Many programmes were using Blackboard for formative assessments, but difficulties in quality assuring Blackboard led to the decision to use Perception for all summative testing. Policy, based on QAA precepts for distributed learning (QAA, 2004), and the BS 7988: 2002 standards (BSI, 2002), and which dovetailed into the existing examinations policies, was formulated. Given the complicated nature of the Perception V3 programme, it was decided that the Learning and Teaching Enhancement Unit would make assessments for staff and a process was initiated to ensure that tests were accurate and fit for purpose. This added to the time taken to create the test. It is envisaged that future tests will be made using the Respondus tool, and delivered to the LTEU in QTI format which should shorten this time.

This assessment exercise produced a number of challenges to the existing risk assessment which formed the basis on the online summative assessment policy:

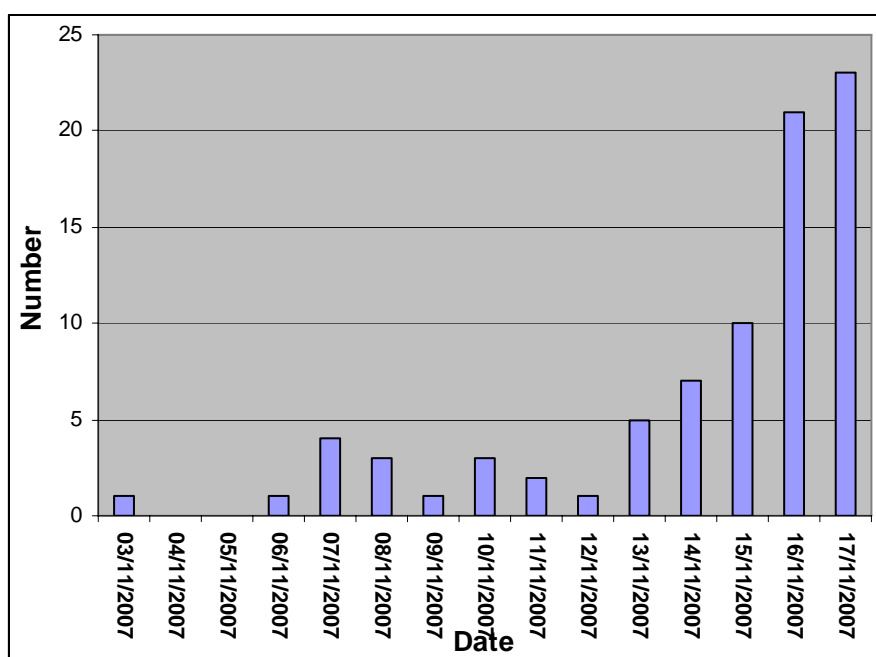
- Students were to be enabled to take the test off campus.
- Students could take the test at any time.
- There was no real way of knowing whether the student who took the test was in fact the correct student.

The practice test was accessed 291 times by 40 students, most students trying it an average of four times before moving to the summative test. Most support issues arose from users off campus. Approximately 33% of users accessed the test at home, and of these 3 were unable to access the test, even after extensive advice on browser settings etc. This indicates a potential problem for students who leave the test until the last minute. 43 students took the test on the last 2 days, 6 leaving it until after core support hours (see Figure 7).

Some students tried to access the summative test before they were ready to take it. This raised the issue of whether they had deliberately looked at the test before taking it. Server logs showed how long each student with a failed attempt had accessed the test, and none had spent more than a few seconds logged in, indicating that they had indeed made a mistake. A decision was made on an individual basis as to whether these students should be able to take the test again.



**Figure 7. Number of tests taken over the period of availability.**



Whereas the risks identified by Zakrzewski and Steven (2000) in the implementation of CAA for summative assessments are mediated via procedures undertaken by academic and support staff, and which were well-understood institutionally and catered for in existing policies, this test raised new aspects of risk which were exacerbated by the students' "last minute" working practices. Students who left the test until the 11<sup>th</sup> hour and who experienced difficulties were dealt with individually, but this is an aspect of testing which takes place over time which needs to be resolved through clear regulations rather than on an ad-hoc basis. The success of this assignment has led to improved communication with the students on the nature of their online assessment.

## **Conclusions and recommendations**

Establishing boundaries regarding staff availability seems key for the success of a CAA exercise. Instructions detailing staff's response time to queries would enhance students' experience, provide a more realistic framework of expectations, and ease the pressure on the staff involved. All the technical problems related to the assessment arose from students accessing the assessment from outside the university network: the problems were not always easy to diagnose, due to differences in Internet providers. It was decided that, in future assessments, students would be encouraged to access the formative test from home, but told to do the summative one from a university computer if they want to have technical support.

Student feedback regarding the computer assessment has been very positive. However, the results should be taken with caution, due to the low number of responses (38.5% of the total). Also, since the questionnaire was voluntary, the respondents are self-selected and not random: only students with some motivation filled it in. On the other hand, it could be said that no student felt so negatively about the process that they needed to fill in the questionnaire. This group was mainly composed of experienced computer users, which may have influenced their opinion of the experience: more data from future cohorts will have to be gathered in order to ascertain the influence of computer experience in students' attitude towards the assessment procedure.

From the lecturer's point of view, the experience was challenging but very positive. The initial time invested preparing the assessment was considerable; however, that time was used in a creative way (designing the tests) as opposed to spent in routine marking. Overall, results indicate the experience was positive for both staff and students. The experience and feedback from this cohort has been very valuable to improve next cohort's assessment; the results for cohort 2006-2007 will be soon available.

### **Acknowledgements**

E. Bertolo would like to thank the staff from the Learning and Teaching Enhancement Unit for the technical support provided, and all the students who answered the evaluation questionnaire.

## References

Adams K., Byers B., Cole R., Ruddick D., Adams D. (2003) Computer-aided Assessment in Chemistry, LTSN Physical Sciences Development Project, <http://www.physsci.heacademy.ac.uk/Resources/DevelopmentProjectsReport.aspx?id=78> (10 January 2007).

BSI (2002) British Standard BS788, Code of Practice for the use of information technology (IT) in the delivery of assessments.

Bullock A. (2003) Using WebCT for computer aided assessment (CAA) - a case study, Teaching Forum, 5, 11-14, <http://www.brookes.ac.uk/virtual/NewTF> (25 May 2006).

Bull J., Danson M. (2004) Computer-assisted Assessment (CAA), LTSN Generic Centre – Assessment Series No 14.

Clarke A., (2001), Designing computer-based learning materials, chapter 4, Gower Publishers Ltd.

Gladwin R. (2005) Getting started with CAA, The HE Academy Physical Sciences Centre.

Jenkins M. (2004) Unfulfilled Promise: formative assessment using computer-aided assessment, Learning and Teaching in Higher Education, 1, 67-80.

King T., Duke-Williams E. (2001) Using Computer-Aided Assessment to Test Higher Level Learning Outcomes, Proceedings of 5th International Computer Assisted Assessment Conference, Univ. of Loughborough, 177-187 <http://www.caaconference.com/pastConferences/2001/proceedings/p1.pdf> (10 January 2006).

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

Overton T. (2003) Key aspects of teaching and learning in experimental sciences and engineering, in A handbook for teaching and learning in higher education, ed. H. Fry, S. Ketteridge, H. Marshall, 2nd Ed, Kogan Page, UK.

Price G. (2006) Computer aided assessment and formative feedback – can we enhance students' early experience at University? HE Academy Physical Sciences Centre Development Project,

<http://www.physsci.heacademy.ac.uk/Resources/DevelopmentProjectsReport.aspx?id=213> (10 January 2006).

QAA (2004) Code of Practice Precept B7 and B8, Assessment of Students. (2004), <http://www.qaa.ac.uk/academicinfrastructure/codeOfPractice/section2/default.asp#assessment> (25 March 2005).

Zakrzewski, S and Steven, C (2000) A Model for Computer-based Assessment: the Catherine wheel principle, *Assessment and Evaluation in Higher Education* 25, 201-215.

**THE FORMATIVE USE OF  
E-ASSESSMENT:  
SOME EARLY IMPLEMENTATIONS,  
AND SUGGESTIONS FOR HOW WE  
MIGHT MOVE ON**

**Andrew Boyle**



# **The Formative Use of e-Assessment: Some Early Implementations, and Suggestions for How We Might Move On**

Andrew Boyle,  
Assessment Research team,  
Regulations and Standards division,  
Qualifications and Curriculum Authority (QCA),  
83 Piccadilly,  
London W1J 8QA.  
0207 509 5349  
BoyleA@qca.org.uk  
<http://www.qca.org.uk/>

## **Abstract**

This paper reviews research into the formative use of e-assessment. The review groups implementations into three areas, and then suggests areas for further research in each area. There are nine areas for further research in total.

The discussion section examines the areas for further research to establish commonalities between them. By this process, it proposes four key issues to inform the future of formative e-assessment research.

The key issues are:

- Better defining those instances where formative e-assessment provides particular benefit over and above benefits that would accrue from the use of formative assessment in any medium.
- Being aware of – and attempting to avoid – formative e-assessment implementations that represent a reduced or impoverished conception of formative assessment.
- Being aware of circumstances in which the introduction of formative e-assessment could lead to increased burdens on classroom practitioners.
- The need to understand how students will be required to adopt novel roles (e.g. different ways of working and communicating) when using formative e-assessment.

## Introduction

Early e-assessment soothsayers made several predictions. An important one was that e-assessment would facilitate a lowering of barriers between assessment and learning. It so happens that the early years of e-assessment implementation have coincided with a heightened interest in formative assessment (FA).

Thus, it is felt timely to conduct a literature review into the formative use of e-assessment. This review looks across studies and attempts to spot frequent implementations of formative e-assessment (eFA), then group and present them to give an insight into what has been done most frequently in this field.

However, this is also a critical review. As well as constructing categories of frequently used implementations of eFA, the review points out issues that are not adequately resolved and suggests further research to rectify omissions or misunderstandings that currently exist. Building upon those suggestions for further research, the review concludes by proposing four key issues for improving the body of research into the formative use of e-assessment (eFA).

## Definitions

### *Formative assessment*

Black & Wiliam (1998a) define formative assessment as follows:

[FA] encompasses all those activities undertaken by teachers, and/or by their students, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged.

Other terms have been used to refer to formative assessment, including 'Assessment for Learning' (AfL) (Black & Wiliam, 1998a) and 'classroom evaluation' (Crooks, 1988).

Formative assessment is often contrasted with summative assessment. Summative assessment is assessment that summarises learning, and which is used for recording and reporting the amount of learning but not for feeding back into learning (Harlen, 2005, p. 208).

### *E-assessment*

E-assessment includes tests that are delivered on-screen, as well as other assessment instruments – in particular e-portfolios. Also, the review encompasses e-learning technologies (such as virtual learning environments – VLEs – and components thereof such as electronic discussion boards, forums and so on).

Cognate terms for e-assessment are included in this review, including: computer-based assessment (CBA) and computer-assisted assessment (CAA). Further, some articles included in the review might not talk about e-



assessment at all. They refer to e-portfolios or e-learning courses, and the use of these technologies for FA purposes.

## Research aims

Thus, the aims of this research are:

- To identify types of implementations that have been used frequently by researchers working in the field.
- Having described key features of implementation types, to suggest ways in which the body of research evidence might be expanded.

In describing eFA implementations and areas for potential further study, an underlying aim is to delineate those areas where eFA provides a distinctive input when compared to formative assessment research more generally<sup>1</sup>.

## Method and scope

This review is inclusive, rather than excluding. It attempts to provide a 'best evidence synthesis' and results that are authentic, faithful and convincing (Black & Wiliam, 2003, p. 629), rather than complying with one or more 'objective' criteria.

This is a thorough review of eFA literature. It is backed up by a selective review of formative assessment literature. It is not a general review of e-assessment<sup>2</sup>.

## Data

109 papers have been considered for this review. Their distribution between eFA and 'plain' FA is shown in the table below:

Formative use of e-assessment	73
Plain formative assessment	25
General policy of e-assessment	11
<b>Total</b>	<b>109</b>

**Table 1: Number of papers in review of different types**

The clear majority of the papers related to eFA. A substantial minority described issues in 'plain' FA research. A third category of 11 papers was also discerned (see, for instance: Bennett, 1998; Bennett, 2002; Wainer,

---

<sup>1</sup> In implementing this 'background aim', regard is had both to those thinkers who suggest that e-assessment will be a necessarily transformative technology (e.g. Bennett, 1998 and 2002), and to more sceptical commentators – who point out how supporters of new technologies have often overstated their potential, and that adoption of such has often led to unexpected consequences (Cuban, 2001).

<sup>2</sup> There are several comprehensive reviews of e-assessment: Ridgway et al, 2004; Sim et al, 2004; Conole & Warburton, 2005.

2000). These papers were early works discussing the potential of e-assessment to transform education; in particular, to facilitate a closer link between assessment and learning.

### **Background findings on formative assessment**

Black and Wiliam (1998b) summarises a fuller description of a comprehensive literature review (Black and Wiliam, 1998a). It poses, and then answers, some questions, including:

- Is there evidence that improving formative assessment raises standards?
- Is there evidence about how to improve formative assessment?

Black and Wiliam (1998b) concludes that there is evidence of substantial learning gains from formative assessment. Further, FA is particularly effective at helping lower-achieving pupils.

Elwood has questioned whether claims for formative assessment's effectiveness in improving learning have been overstated. She suggests that learning gains may be partly accounted for by error variance in test scores, and that gains of learners in FA studies may result from sources other than the FA intervention (Elwood, 2006, p. 227).

Black and Wiliam (1998b) describe how to improve FA practice:

- Feedback to any pupil should be about the particular qualities of his or her work, with advice on what he or she can do to improve, and should avoid comparisons with other pupils.
- For formative assessment to be productive, pupils should be trained in self-assessment so that they can understand the main purposes of their learning and thereby grasp what they need to do to achieve.
- Opportunities for pupils to express their understanding should be designed into any piece of teaching, for this will initiate the type of interaction in which formative assessment aids learning.
- The dialogue between pupils and a teacher should be thoughtful, reflective, focused to evoke and explore understanding, and conducted so that all pupils have an opportunity to think and to express their ideas.

Thus, formative assessment has several aspects – concerning the nature of classroom interactions between teachers and learners (including the way that questions are asked and answered), peer- and self-assessment and the nature of written feedback.

Feedback is a central issue in FA (Sadler, 1989; Sadler, 1998). This includes both the way that teachers interact with pupils in speech, and the nature of written feedback. Written comments are more effective when they are

specific (e.g. not just saying 'an excellent effort') and when they permit a pupil to 'close the gap' between current and desired performance.

There is controversy as to whether written feedback should contain a mark or grade. Black and Wiliam (1998a; 1998b) state that written comments should not contain a mark or grade. Effectively implemented 'comment-only' marking is more likely to give pupils the necessary information to close the learning gap, whereas recipients are more likely to focus on marks or grades at an emotive level (as a comment on their personal worth) rather than as providing a spur to improve work.

Smith and Gorard (2005) cautiously reported an implementation of comment-only marking that did not work as Black and Wiliam would have predicted<sup>3</sup>. In Smith and Gorard's small study, pupils receiving comment-only feedback made inferior progress to that of other classes.

Most FA research has been about a range of classroom practices rather than evaluating assessment instruments and questions. However, Wiliam (2005) proposed that good FA items might have the following properties, which are different to those for good summative assessment items:

- Can have more (or less) than one correct answer
- Items need to be generative
  - of learning
  - of insights into learning
  - of insights into how to promote learning
- Distractors must be explicitly connected to incorrect or incomplete conceptions (facets)
- Item responses must provide clues to effective action

Thus, FA research has examined an area of interest in some depth, and has established some fairly clear principles. There are some reservations about the extent to which reported gains represent genuine effects and a feeling that there needs to be a deeper understanding of the effects of error variance in assessment scores; this is quite a common concern in assessment research. Also, the ways in which clear principles are interpreted when rolled out across an educational sector remain worthy of further study.

These two *caveats* are worth bearing in mind when considering eFA research.

---

<sup>3</sup> Black et al (2005) attempted to rebut Smith and Gorard's tentative findings.

## **Review findings: eFA implementations and areas for further research**

In the following part of the paper, common implementations of eFA are presented and elaborated. Then, they are critiqued and suggestions for further research are made.

### **Finding 1**

**Electronic technologies provide a range of new tools that classroom teachers can use to create formative assessments to suit their and their students' needs.**

Many eFA implementations used different task or item types and varied assessment designs. These include:

- Variations on the theme of multiple-choice questions (MCQs):
  - 'formative quizzes' (Cassady & Grindley, 2005)
  - MCQ-based tests made available to students for frequent use (Baggott & Rayne, 2001; Peat & Franklin, 2002)
  - MCQ tests provided for students to allow them to practise the format of the final exam (Cassady et al, 2001; Peat et al, 2005) or as revision (Irving et al, 2000)
  - MCQs adapted to allow students to indicate how confident they are in a particular answer before giving it (Farrell et al, 2005; Gardner-Medwin & Gahan, 2003)
- More advanced or 'sophisticated' (Boyle, 2005) e-assessment tasks – including those rich in interactivity and multimedia:
  - Scenario-based assessments (Crisp & Ward, 2005)
  - Simulation-based assessments (Young & Cafferty, 2003)
  - Concept maps used for formative assessment of collaborative problem solving (Hsieh & O'Neill, 2002)
- Test designs that are specific to e-assessment<sup>4</sup>:
  - Computer Adaptive Testing (Lilley et al, 2004; Lilley et al, 2005; Yong & Higgins, 2004)
- The use of e-portfolios to facilitate closely integrated formative and summative assessment (McGuire et al, 2004; McGuire, 2005; Woodward & Nanlohy, 2004).

---

<sup>4</sup> Or at least can be done much more efficiently electronically.

- The use of communications tools such as electronic discussion boards and forums for self- and peer feedback in e-learning courses (Keppell, & Carless, 2006; Keppell et al, 2006; Lin et al, 2001).

These examples perhaps support Bennett's (2001) contention that e-assessment will give rise to mass customisation of assessment products; that is, the ability of educational practitioners to use technologies to provide assessment solutions to suit their particular teaching and learning needs.

However, the early usage of e-assessment instruments for formative purposes also gives rise to areas in need of clarification. These are set out below.

#### Finding 1: Area for further research (a)

*'Plain' FA research has suggested that formative and summative test questions may have different ideal characteristics. However, existing implementations of eFA have tended to take item and task types that originate from summative assessment. New research should attempt to establish the ideal characteristic of items and tasks used for eFA.*

'Plain' FA research has not focused much on the nature of test instruments used. eFA provides a range of instruments that practitioners may find useful. However, many early implementations have simply applied summative test and question designs to the formative arena. This may be appropriate, but an interesting new strand of research might build upon William's contrasting of different properties of good formative and summative items and suggest distinctive features of good eFA items.

#### Finding 1: Area for further research (b)

*eFA implementations have not sufficiently distinguished notions of 'formative assessment' from 'exam revision' or 'becoming acquainted with summative test formats'. Future research should make that distinction more clearly.*

The body of FA knowledge has a range of facets. However, several of the eFA papers equate exam revision or practice testing with FA. This is not to say that exam revision is a bad thing; it has a role to play in decreasing students' test anxieties (Cassady & Gridley, 2005) and frequent use of e-assessment quizzes can help students learning from distance to remain motivated and focused (Baggott & Rayne, 2004). Nonetheless, the danger of equating eFA with exam revision is that it will represent a reduced notion when compared to the complete body of formative assessment research.

#### Finding 1: Area for further research (c)

*Early implementations of eFA tended to involve innovators developing their own questions. Further research should investigate whether it is realistic for all teachers to write test questions for eFA or whether – if*

*teachers merely select from a bank of questions – anything is lost by that process.*

Early enthusiasts have developed eFA systems by writing their own questions. It is debatable whether the wider body of teachers would have the necessary time, motivation and skills to write large numbers of high quality test questions.

If teachers using eFA do not write their own questions, an alternative might be for them to use products that contain pre-written questions. Further research might fruitfully investigate the implications of using such eFA products. For example, would the use of a pre-written bank decrease a teacher's ability to tailor questions to suit the needs of learners in her own class?

## Finding 2

**e-assessment functionality permits formative feedback to be given in a variety of ways that is not possible in 'plain' FA.**

Developers of eFA systems have found a range of ways to deliver formative feedback, including:

- Formative feedback given differentially for entirely correct, partially correct and entirely incorrect answers (Wood and Burrow, 2002)
- Feedback as references to textbook chapters (Buchanan, 2000)
- Feedback realised as rich multimedia (Mackenzie, 2000)
- Feedback as references to web sites (Mackenzie, 2003; Clarke et al, 2004)
- Feedback delivered within questions (CIAD, 2005) after each question, or at the end of each timed session (Baggott and Rayne, 2001)
- Rich-media feedback as a stimulus to peer-to-peer discussion of content (Mackenzie, 2003)

Advocates of e-portfolio systems have suggested several advantages that can accrue when e-portfolios are used to provide feedback. These include:

- e-portfolio authoring encourages teachers and students to view drafts of work, and interact about them. The process of generating work is forefronted, rather than merely concentrating on the final product (Twining et al, 2006, p. 55).
- Tools in e-portfolios can allow teachers to ask students to upload materials at significant stages, thus illustrating what the students believe to be progress (an important element of self-assessment) (Twining et al, *ibid.*; McGuire et al, 2004, p. 4).
- Communications tools associated with e-portfolios can allow for the provision of varied feedback with respect to: authors (fellow students or teachers), formality, and mode of communication – writing or speech (McGuire, 2005, p. 267). Such variety can be useful for facilitating

learning by students of different dispositions, experiences and cognitive styles.

Researchers have reported their uses of online e-learning technologies – often on distance learning courses. They have described how communications technologies (such as message boards and discussion forums) have allowed them to provide innovative feedback to assist learning, including:

- Students taking part in online discussions, and being required to submit a specified number of contributions (Goodfellow & Lea, 2005) – a form of peer feedback
- Students keeping a reflective journal (Keppell & Carless, 2006) – feedback to oneself or self-assessment
- Students rating peers' work quite formally – including giving marks (Bhalerao & Ward, 2001) or less formally (Lin et al, 2001), including taking part in collaborative group activities (MacDonald, 2004)

Thus, practitioners have used a range of e-assessment technologies to provide feedback to students. However, there also remain questions arising from these implementations, which may allow researchers to theorise the use of eFA to provide feedback more comprehensively.

#### Finding 2: Area for further research (a)

*Where teachers use extensions to e-test delivery systems to provide feedback to students, further research should establish principles for the design of such feedback so as to optimise students' learning opportunities.*

Several researchers have attempted to systematise understanding of the qualities of effective feedback when using e-tests (e.g. Hanson et al, 2001; Hsieh and O'Neill, 2002; Clarke et al 2004; Brettell et al, 2005). However, questions remain to be resolved, including:

- Does the stricture from 'plain' FA that feedback should be made up of comments but not grades apply when e-tests are used? If so, does this disable one of the most obvious uses of an e-assessment system for formative purposes?
- To what extent is engagement with rich media or interactive feedback synonymous with deep learning? Or are there circumstances where varied media or interactive possibilities distract learners and lead to superficiality (e.g. clicking through links without truly processing the content of web pages – see Clarke et al, 2004)?

- Is the impact of feedback related to students' learning styles? For example, the work of Brettell et al (2005) to distinguish responses to feedback of 'deep', 'strategic' and 'surface apathetic' learners could profitably be extended.

#### Finding 2: Area for further research (b)

*Where e-portfolios are used with the aim of facilitating the giving of feedback (teacher-to-student; student-to-other-student and student-to-self), logistical or ergonomic studies should be conducted to make sure that users find it practical to give feedback via the portfolio tools.*

McGuire (2005) noted that e-portfolios were not 'an easy option', but asserted that they were worthwhile in that they allowed the giving of rich feedback. It will be important to ensure that this potential is not lost; teachers can find it burdensome to provide comments of sufficient quality on students' work (Smith & Gorard, 2005). ICT elements of portfolios should reduce this burden, and thus facilitate the giving of high-quality feedback.

#### Finding 2: Area for further research (c)

*Where online tools such as discussion boards and electronic forums are used to facilitate feedback, research should investigate the impact of cultural factors on students' ability to give peer feedback.*

Students giving feedback via electronic tools may suffer if they do not understand cultural norms relating to the giving of feedback. This may have two facets; many online distance learning courses will involve students from different parts of the world. Such students may have differing prior assumptions about commenting on colleagues' work. This may be accentuated when they are working remotely and thus have fewer opportunities to interact face-to-face with peers and/or teachers.

Misunderstanding cultural norms can occur when students are from different countries. However, it can also occur when students have not internalised the norms associated with academic discourses. In particular, early thinking on electronic communication asserted that new communication forms blurred the boundaries between writing and speech – e.g. writing with reduced formality and increased interactivity would be more like speech (Lawler & Dry, 1998). However, giving written feedback on peers' work in an electronic environment is a novel discourse form, and its relationship to formal academic writing remains to be established (Russell et al, 2006). Further research could set out similarities and differences in these two ways of writing and help students to effectively switch between the two.



### Finding 3

**eFA applications can be used remotely in time (asynchronously). This facility of electronic tools provides a resource which is not easily replicated via pencil-and-paper materials.**

Some papers in the review present implementations in which students have been able to go away and use formative assessment materials. Many of the reported studies involved Higher Education classes – often those with new undergraduates. The asynchronicity afforded by electronic materials was said to have the following advantages:

- The use of remote self-access formative assessment materials was associated with reduced examination stress (Baggott and Rayne, 2001; Cassady et al, 2001; Cassady & Gridley, 2005).
- The eFA materials were popular with students and motivating (Blayney & Freeman, 2003).
- The provision of eFA materials freed up teachers' time and thus facilitated courses with high student:teacher ratios (Peat et al, 2005).
- The use of self-assessment eFA materials allowed students to increase their self-regulation (Brettell et al, 2005), in particular to get used to learning independently in tertiary study (Peat et al, 2005).
- The asynchronous aspect of online discussions, added to the fact that evidence of discussion content could be reviewed (e.g. by looking at 'threads' of groups on a web site), facilitated participants' enhanced reflection (Russell et al, 2005).

However, some researchers have noted areas that require clarification.

- There appears to be some relationship between learners' cognitive styles and or their motivations and their use of electronic self-assessment materials. In particular, those who are already skilled in self-regulation may get more benefit from the materials than those who are not (Lin et al, 2001). Also, usage patterns may differ between those learners who are intrinsically interested in learning for its own sake and 'pragmatists' (Keppell & Carless, 2006).
- There are varying results with respect to usage patterns of asynchronous eFA materials. Some researchers report that students used the materials throughout their courses (Bryan et al, 2005), whilst others found usage was concentrated in the period running up to the summative assessment (Pitt & Gunn, 2004).

In addition to those reservations about the corpus of research evidence on the asynchronous use of eFA materials, the current review adds two further areas that should be clarified so that research evidence is more complete.

### Finding 3: Area for further research (a)

*Although several studies have claimed that use of eFA materials is associated with learning gains, the bases on which they do so are generally not well founded. If a claim is to be made that eFA provides enhanced learning gains over and above 'plain' FA, then better designed studies need to be conducted.*

A substantial number of the eFA papers in this review (especially those that reported on the asynchronous/self-access use of formative materials) claimed that students who used the materials had an attainment benefit. However, in almost all cases these claims were undermined by an aspect of the research design. For instance, studies were conducted with small cohorts, or the difficulty of two years' tests was not properly equated or studies confounded variables (e.g. did the students using eFA score more highly because it was an eFA intervention, or did they score more highly because they worked harder?).

Thus, an important claim of the eFA literature has not been robustly established. That 'plain' FA is associated with learning gains is an important tenet of that literature, but it might be interesting for researchers to design studies that build from the work of plain FA researchers and show particular ways in which eFA supports enhanced attainment.

### Finding 3: Area for further research (b)

*The equating of eFA with self-assessment is strongly associated with patterns of learning in tertiary education. It would interesting to see whether the self-access paradigm could be imported into secondary or primary education.*

The literature reporting the asynchronous use of eFA materials is strongly associated with tertiary education. Taking online quizzes and the like is seen as a way to encourage new undergraduates to manage their study in an environment where they were expected to take more responsibility than at school.

It would be useful to see what issues would crop up if e-self-assessment materials were widely used by school-age students. For example, school teachers may feel a greater obligation to moderate feedback (e.g. to avoid students receiving potentially demotivating critical feedback). Other issues not apparent in the tertiary sector might also arise (e.g. the role of parents in supporting their children's online learning).

## Discussion

The aim of this review was to map implementations of eFA, and to suggest areas for further research. In doing so, the intention was also to describe those areas where the use of e-assessment for formative purposes provided a distinctive contribution; different to anything that came from the wider body of formative assessment research.

Starting from implementation has the virtue of being a 'reality check'; giving an overview of the state of the art at a particular point. It affords the possibility of description of actual practice. Evaluation of that practice can then suggest the extent to which implementations have fulfilled aspirations for eFA. It can also facilitate a re-focusing on areas that need increased attention; especially if such areas are unexpected.

However, it may be that working from implementations can give a somewhat fragmented picture of the unique features of formative e-assessment. For that reason, attention has been paid to the nine 'areas for further research' that have been proposed in this review. These have been examined to search for commonality between them.

In fact, there does appear to be some commonality between the nine areas for further research, and so it is possible to propose four 'super categories' or key issues that might guide future eFA research.

### Key issue 1

**eFA research needs to better define the ways in which the electronic element provides added benefit above and beyond 'plain' FA use.**

This key issue requires thinking about eFA to demonstrate its added value beyond plain FA. Also, however, it would critique eFA implementations that simply adopted summative e-assessment designs without showing their suitability for the formative purposes. The key issue arises from the following areas for further research:

- 1a: use of e-assessment instruments by practitioners
- 2a: provision of feedback from e-assessment instruments
- 3a: need for better-designed studies to demonstrate attainment benefits

### Key issue 2

**Those promoting eFA implementations should ensure that eFA does not amount to a reduced or impoverished notion when compared to the full understanding of formative assessment.**

This key issue is – in some senses – the converse of the first. However, it goes somewhat further; whilst key issue 1 imposes a positive duty on eFA to show distinctive benefit, this key issue notes the possibility that eFA can have negative consequences. It arises from the following areas for further research:

- 1b: equating FA with exam revision
- 2a: provision of feedback from e-assessment instruments

### Key issue 3

**Attention should be given to the danger that eFA might impose new burdens on teachers (and – to some extent – students).**

ICT innovations are often touted as labour saving. However, if they are not well designed (or specifically fit for an educational purpose, Cuban (2001, p. 170)), they may not be as widely adopted as expected.

This key issue arises from the following areas for further research:

- 1c: requirement for teachers to write their own test questions
- 2b: need for e-portfolios to provide manageable systems for giving feedback

### Key issue 4

**Students using eFA applications will sometimes be required to take on novel roles. The ways in which students adapt to such novel roles should be monitored.**

Students may need to work more independently than previously, or to communicate according to cultural or social norms which are alien to them. The extent to which they are successful in so adapting could be an important area of eFA research.

This key issue arises from the following areas for further research:

- 2c: cultural factors in the use of electronic communications tools
- 3b: strong element of independent working and self-assessment in eFA

Different sets of key issues may be arguable, but it is proposed that if eFA research were to focus on these four areas, then it would be stronger, and have a chance of leading to more principled implementations.

## Bibliography

Baggott, G. & Rayne, R. (2001) *Learning support for mature, part-time, evening students: providing feedback via frequent, computer-based assessments* in Danson, M. (ed.) Fifth International Computer Assisted Assessment (CAA) Conference Proceedings, Loughborough University, July 2001. <http://www.caaconference.com/>.

Baggott, G. & Rayne, R. (2004) *Student perceptions of computer-based formative assessments in a semi-distance module* in Ashby, M. (ed.) Eighth International Computer Assisted Assessment (CAA) Conference Proceedings, Loughborough University, 6th and 7th July 2004. <http://www.caaconference.com/>.

Bennett, R.E. (1998) *Reinventing Assessment: speculations on the future of large-scale educational testing*. <ftp://ftp.ets.org/pub/res/reinvent.pdf>.

Bennett, R.E. (2001) *How the internet will help large-scale assessment reinvent itself*. Education Policy Analysis Archives, Volume 9 Number 5. <http://epaa.asu.edu/epaa/v9n5.html>.

Bennett, R.E. (2002) *Inexorable and Inevitable: The Continuing Story of Technology and Assessment*. The Journal of Technology, Learning and Assessment (JTLA), Volume 1, Number 1.

Bhalerao, A. & Ward, A. (2001) *Towards electronically assisted peer assessment: a case study*. Association for Learning Technology Journal (ALT-J) 9(1) pp. 26 – 37.

Black, P. & Wiliam, D. (1998a) *Assessment and classroom learning*. Assessment in Education: principles, policy & practice, 5(1), pp. 7 – 73.

Black, P. & Wiliam, D. (1998b) *Inside the Black Box: Raising Standards Through Classroom Assessment*. (London: King's College London School of Education).

Black, P. & Wiliam, D. (2003) *'In praise of educational research': formative assessment*. British Educational Research Journal, Volume 29, Number 5, pp. 623 – 637.

Black, P., Harrison, C., Hodgen, J., Marshall, B. and Wiliam, D. (2005) *The dissemination of formative assessment: a lesson from or about, evaluation*. Research Intelligence, 2005, p. 12.

Blayney, P. & Freeman, M. (2003) *Automated marking of individualised spreadsheet assignments: the impact of different formative self-assessment options* in Christie, J. (ed) Seventh International Computer Assisted Assessment (CAA) Conference Proceedings, Loughborough University, July 2003. <http://www.caaconference.com/>.

Boyle, A. (2005) *Sophisticated Tasks in E-Assessment: What are they? And what are their benefits?* in Danson, M. (ed) Ninth International Computer Assisted Assessment (CAA) Conference Proceedings, Loughborough University, 5th and 6th July 2005. <http://www.caaconference.com/>.

Brettell, S., Durham, J. & McHanwell, S. (2005) *'Well nobody reads learning outcomes do they?' – An evaluation of CAA and its feedback on directed student learning* in Danson, M. (ed) Ninth International Computer Assisted Assessment (CAA) Conference Proceedings, Loughborough University, 5th and 6th July 2005. <http://www.caaconference.com/>

Bryan, N. & Glasfurd-Brown, G. (2005) The SPRinTA project: supporting student assessment through a portal in Danson, M. (ed) Ninth International Computer Assisted Assessment (CAA) Conference Proceedings, Loughborough University, 5th and 6th July 2005. <http://www.caaconference.com/>.

Buchanan, T. (2000) *The efficacy of a World-Wide Web mediated formative assessment*. Journal of Computer Assisted Learning Volume 16 Issue 3, p. 193.

Cassady, J.C. & Gridley, B.E. (2005) *The Effects of Online Formative and Summative Assessment on Test Anxiety and Performance*. Journal of Technology, Learning and Assessment (JTLA), Volume 4, Number 1, October 2005.

Cassady, J.C., Budenz-Anders, J., Pavlechko, G. & Mock, W. (2001) *The effects of internet-based formative and summative assessment on test anxiety, perceptions of threat, and achievement*. Paper presented at the Annual meeting of the American Educational Research Association (AERA) (Seattle, Wa., April 10 - 14 2001).

Centre for Interactive Assessment Development (CIAD) (2005). *TRIADS Functionality for Formative Assessment*. <http://www.derby.ac.uk/ciad/formative.php>.

Charman, D. & Elmes, A. (1998). *A computer-based formative assessment strategy for a basic statistics module in geography*. Journal of Geography in Higher Education, 22(3), pp. 381-385.

Clarke, S. Lindsay, K., McKenna, C. & New, S. (2004) *INQUIRE: a case study in evaluating the potential of online MCQ tests in a discursive subject*. Association for Learning Technology Journal (ALT-J), Volume 12, Number 3, September, pp. 249 – 260.

Condie, R. & Munro, B. (2007) The impact of ICT in schools – a landscape review. <http://publications.becta.org.uk/display.cfm?resID=28221&page=1835>.

Conole, G. & Warburton, B. (2005) *A review of computer-assisted assessment*. Association for Learning Technology Journal (ALT-J), Volume 13, Number 1, March, pp. 17-31.

Crisp, V. & Ward, C. (2005) *The PePCAA project: formative scenario-based CAA in psychology for teachers* in Danson, M. (ed) Ninth International Computer Assisted Assessment (CAA) Conference Proceedings, Loughborough University, 5th and 6th July 2005. <http://www.caaconference.com/>.

Crooks, T.J. (1988) *The impact of classroom evaluation practices on students*. Review of Educational Research, 58, pp. 438 – 481.

Cuban, L. (2001) *Oversold and underused: computers in the classroom*. Cambridge, Ma.: Harvard University Press.

Elwood, J. (2006) *Formative assessment: possibilities, boundaries and limitations*. Assessment in Education: Principles, Policy & Practice, Vol. 13, No. 2, July 2006, pp. 215–232.

Farrell, G., Farrell, V. & Leung, Y.K. (2005) *A comparison of Blackboard CAA and an innovative self-assessment tool for formative assessment* in Danson, M. (ed) Ninth International Computer Assisted Assessment (CAA) Conference Proceedings, Loughborough University, 5th and 6th July 2005. <http://www.caaconference.com/>.

Gardner-Medwin, A.R. & Gahan M. (2003) *Formative and summative confidence-based assessment* in Christie, J. (ed) Seventh International Computer Assisted Assessment (CAA) Conference Proceedings, Loughborough University, July 2003. <http://www.caaconference.com/>.

Goodfellow, R. & Lea, M.R. (2005) *Supporting writing for assessment in online learning*. Assessment & Evaluation in Higher Education, Volume 30, Number 3 / June 2005, pp. 261 – 271.

Hanson, J., Millington, C. & Freewood, M. (2001) *Developing a methodology for online feedback and assessment* in Danson, M. (ed.) Fifth International Computer Assisted Assessment (CAA) Conference Proceedings, Loughborough University, July 2001. <http://www.caaconference.com/>.

Harlen, W. (2005) *Teachers' summative practices and assessment for learning – tensions and synergies*. The Curriculum Journal, Vol. 16, No. 2, June 2005, pp. 207 – 223.

Hsieh, I.L.G. & O'Neil, H.F. Jr. (2002) *Types of feedback in a computer-based collaborative problem-solving group task*. Computers in Human Behavior, Volume 18, Issue 6, November, Pages 699 – 715.

Irving, A., Read, M., Hunt, A. & Knight, S. (2000) *Use of information technology in exam revision* in Danson, M. (ed.) Fourth International

Computer Assisted Assessment (CAA) Conference Proceedings, Loughborough University, July 2000. <http://www.caaconference.com/>.

Keppell, M. & Carless, D. (2006) *Learning-oriented assessment: a technology-based case study*. Assessment in Education Vol. 13, No. 2, July 2006, pp. 179–191.

Keppell, M., Au, E. Ma. A. & Chan, C. (2006) *Peer learning and learning-oriented assessment in technology-enhanced environments*. Assessment & Evaluation in Higher Education, Volume 31, Number 4 / August 2006, pp. 453 – 464.

Lawler, J. & Dry, H.A. (2008) *Using computers in linguistics: a practical guide*. London: Routledge.

Lilley, M. Barker, T. & Britton, C. (2004) *The generation of automated student feedback for a Computer-Adaptive Test* in Ashby, M. (ed.) Eighth International Computer Assisted Assessment (CAA) Conference Proceedings, Loughborough University, 6th and 7th July 2004. <http://www.caaconference.com/>.

Lilley, M. Barker, T. & Britton, C. (2005) *Automated feedback for a Computer-Adaptive Test: a case study* in Danson, M. (ed) Ninth International Computer Assisted Assessment (CAA) Conference Proceedings, Loughborough University, 5th and 6th July 2005. <http://www.caaconference.com/>.

Lin, S.S.J., Liu, E.Z.F. & Yuan, S.M. (2001) *Web-based peer assessment: feedback for students with various thinking-styles*. Journal of Computer Assisted Learning, Volume 17 Issue 4 p. 420.

Macdonald, J. (2004) *Developing competent e-learners: the role of assessment*. Assessment & Evaluation in Higher Education, Volume 29, Number 2 / April 2004, pp. 215 – 226.

Mackenzie, D. (2000) Production and delivery of TRIADS Assessments on a university-wide basis in Danson, M. (ed.) Fourth International Computer Assisted Assessment (CAA) Conference Proceedings, Loughborough University, July 2000. <http://www.caaconference.com/>.

Mackenzie, D. (2003) *Assessment for E-learning: what are the features of an ideal E-assessment system?* in Christie, J. (ed) Seventh International Computer Assisted Assessment (CAA) Conference Proceedings, Loughborough University, July 2003. <http://www.caaconference.com/>.

McGuire, L. (2005) *Assessment using new technology*. Innovations in Education and Teaching International, Volume 42, Number 3, pp. 265 – 276.

McGuire, L., Roberts, G. & Moss, M. (2004) *Final report to QCA on the eVIVA project*. <http://210.48.101.74/images/Eviva%20Final%20Report.pdf>.



Natriello, G. (1987) *The impact of evaluation processes on students*. Educational Psychologist, 22, pp. 155 – 175.

Peat, M. & Franklin, S. (2002) *Supporting student learning: the use of computer-based formative assessment modules*. British Journal of Educational Technology, v.33 n5, p. 515 – 523.

Peat, M., Franklin, S., Devlin, M. & Charles, M. (2005) *Revisiting the impact of formative assessment opportunities on student learning*. Australasian Journal of Educational Technology, 21(1), 102 – 117.

Pitt, S.J. & Gunn, A. (2004) *The value of computer based formative assessment in undergraduate biological science teaching*. Bioscience Education e-Journal, Volume 3: May 2004.

Prins, F.J., Sluijsmans, D.M.A., Kirschner, P.A. & Strijbos, J-W. (2005) *Formative peer assessment in a CSCL environment: a case study*. Assessment & Evaluation in Higher Education, Volume 30, Number 4 / August 2005, pp. 417 – 444.

Ridgway, J., McCusker, S. and Pead, D. (2004) *Literature review of e-assessment*. [http://www.nestafuturelab.org/research/reviews/10\\_01.htm](http://www.nestafuturelab.org/research/reviews/10_01.htm).

Russell, J., Elton, L., Swinglehurst, D. & Greenhalgh, T. (2006) *Using the online environment in assessment for learning: a case-study of a web-based course in primary care*. Assessment & Evaluation in Higher Education, Volume 31, Number 4 / August 2006, pp. 465 – 478.

Sadler, D.R. (1989) *Formative assessment and the design of instructional systems*. Instructional Science, 18, pp. 119 – 144.

Sadler, D.R. (1998) *Formative assessment: revisiting the territory*. Assessment in Education: Principles, Policy & Practice, 5(1), pp. 77 – 84.

Sim, G., Holifield, P. and Brown, M. (2004) *Implementation of computer assisted assessment: lessons from the literature*. Association for Learning Technology-Journal (ALT-J), 12 (3) 215 – 229.

Smith, E & Gorard S. (2005) *'They don't give us our marks': the role of formative feedback in student progress*. Assessment in Education: Principles, Policy & Practice, Volume 12, Number 1, pp. 21 – 38.

Twining, P., Broadie, R., Cook, D., Ford, K. Morris, D. Twiner, A. & Underwood, J. (2006) *Educational change and ICT: an exploration of Priorities 2 and 3 of the DfES e-strategy in schools and colleges (The current landscape and implementation issues)*. [http://partners.becta.org.uk/page\\_documents/research/educational\\_change\\_and\\_ict.pdf](http://partners.becta.org.uk/page_documents/research/educational_change_and_ict.pdf).

Wainer, H. (2000) *CATs: whither and whence*. Psicológica 21, pp. 121 – 133.

William, D. & Thompson, M. (2006) *Integrating assessment with learning: what will it take to make it work?* in Dwyer, C. A. (ed) *The future of assessment: shaping teaching and learning*. Mahwah, NJ: Lawrence Erlbaum Associates.

William, D. (2005) *Formative assessment and the regulation of learning*. Paper presented at UC Berkeley seminar, March 2005, Berkeley, CA.

Wood, J. & Burrow, M. (2002) *Formative Assessment in Engineering Using 'TRIADS' Software* in Danson, M. (ed) *Sixth International Computer Assisted Assessment (CAA) Conference Proceedings*, Loughborough University, 5th and 6th July 2005. <http://www.caaconference.com/>.

Woodward, H. & Nanlohy, P. (2004) *Digital portfolios: fact or fashion?* *Assessment & Evaluation in Higher Education*, Volume 29, Number 2 / April 2004, pp. 227 – 238.

Yong, C-F. & Higgins, C.A. (2004) *Self-assessing with adaptive exercises* in Ashby, M. (ed.) *Eighth International Computer Assisted Assessment (CAA) Conference Proceedings*, Loughborough University, 6th and 7th July 2004. <http://www.caaconference.com/>.

Young, A. & Cafferty, S. (2003) *Simulation as a tool for computer-assisted formative assessment: First aid as a case study* in Christie, J. (ed) *Seventh International Computer Assisted Assessment (CAA) Conference Proceedings*, Loughborough University, July 2003. <http://www.caaconference.com/>.

*All web links were live on 20 February 2007.*

## **Appendix 1: Sources consulted in research**

### *Research databases and specialist search engines*

#### Research databases

- Education Resources Information Center (ERIC)  
(<http://www.eric.ed.gov/>)
- Bibliography on Computer Based Assessment and Distance Learning  
(<http://iinwww.ira.uka.de/bibliography/Misc/cba.html>)
- EBSCO Host Academic Search premier

#### Specialist search engines

- <http://scholar.google.com/>
- <http://www.scirus.com>
- <http://citeseer.ist.psu.edu/>

### *Journals*

#### Comprehensively handsearched journals

- Assessment in Education: Principles, Policy & Practice
- Assessment & Evaluation in Higher Education
- Journal of Computer Assisted Learning (JCAL)
- Association for Learning Technology Journal (ALT-J)
- Journal of Technology, Learning and Assessment (JTLA)
- British Journal of Educational Technology (BJET)
- Research Papers in Education
- British Educational Research Journal (BERJ)
- Curriculum Journal

#### Other journals that provided articles for this project include

- Australasian Journal of Educational Technology
- Bioscience Education e-Journal
- CAL-elaborate
- Cambridge Journal of Education
- Computers in Human Behavior
- Educational Psychologist
- Engineering Education
- Innovations in Education and Teaching International
- Innovations in Education and Training International
- Journal of Dental Education

- Journal of Educational Multimedia and Hypermedia
- Journal of Geography in Higher Education
- Learning and Teaching in Higher Education
- Measurement
- Psicológica
- Research Intelligence
- Review of Educational Research
- Studies in Continuing Education
- Teaching Mathematics and Its Applications
- The Internet and Higher Education

#### Conference archives

- Computer-assisted Assessment (CAA) conference (<http://www.caaconference.com/>)
- Association for Educational Assessment – Europe (<http://www.aea-europe.net/>)
- International Association for the Evaluation of Educational Achievement (IAEA) ([http://www.iaea.info/index.php?option=com\\_conferences&Itemid=45](http://www.iaea.info/index.php?option=com_conferences&Itemid=45))

Other conferences provided articles for the research, but they did not have comprehensive central archives of papers.

#### *Other sources of information*

Review articles that provided references

- Ridgway et al (2004)
- Conole and Warburton (2004)
- Sim et al (2005)

Lists of ‘key ‘plain formative assessment’ papers’ provided by:

- Bill Boyle, Centre for Formative Assessment Studies (CFAS), University of Manchester
- Paul Newton, Qualifications and Curriculum Authority.

**PRINCIPLES FOR THE  
REGULATION OF E-ASSESSMENT  
AN UPDATE ON DEVELOPMENTS**

**Andrew Boyle**



# Principles for the Regulation of e-Assessment

## An Update on Developments

Andrew Boyle,  
Assessment Research team,  
Regulations and Standards division,  
Qualifications and Curriculum Authority (QCA),  
83 Piccadilly,  
London W1J 8QA.  
0207 509 5349  
BoyleA@qca.org.uk  
<http://www.qca.org.uk/>

### Abstract

The Qualifications and Curriculum Authority (QCA) is a statutory body in England, sponsored by the Department for Education and Skills (DfES). Its functions are set out in the 1997 Education Act, and subsequent amendments. QCA maintains and develops the national curriculum and associated assessments, tests and examinations, and regulates qualifications offered in schools, colleges and workplaces. Its regulatory role covers all qualifications except those awarded by higher education institutions. QCA's role is restricted to England, although it regulates qualifications jointly with its regulatory partners in Wales and Northern Ireland, and works closely with its counterpart in Scotland.

In furtherance of its regulatory role, the QCA (and its sister regulators in Wales and Northern Ireland) has published a set of *regulatory principles for e-assessment*. This presentation will describe background issues that have an impact on e-regulation, the thinking that motivated the development of the *principles*, and report findings from a public consultation on the *principles*, and initial research into the regulation of e-assessment.

Several background factors potentially impact on how e-assessment may be regulated. These include:

- The history of the regulation of qualifications in England  
Many current concerns about assessment standards and integrity have parallels going back to the beginning of large-scale examinations in England. This has, historically, affected the balance that has been struck between protecting the public interest and facilitating providers of qualifications.

- Changes in industrial organisation have meant that old-style regulation is no longer viable

Information and Communications Technologies (ICTs) have fundamentally affected the ways in which industrial activity is organised and conducted. This, in turn, has a profound impact on the way in which regulation can function. For example, where there were once separate regulators for broadcasting, Internet content and telephone communication, confluence of these media channels requires a new approach to regulation.

- Cultural differences can be observed in approaches to regulation of ICT-influenced industries.

Specifically with respect to the Internet, European jurisdictions have tended to emphasise the maintenance of public confidence (and therefore have adopted more proactive regulation), whereas the US has perceived freedom of expression as the main benefit and therefore has had a more relaxed attitude to Internet regulation.

- Socio-legal scholars have described an increase in the use of non-traditional methods for dispute resolution and governance.

Systems for dealing with issues which might previously have been resolved by recourse to formal legal mechanisms have been observed to change. For example, there appears to be a wider use of facilitative, flexible and subtle techniques. Such techniques borrow from the private sector, often depend on self- or peer-reporting and have been shown to be more effective than traditional approaches to delivering policy objectives.

‘Soft-law’ approaches do have associated problems – including their appropriateness for immature markets and how to integrate novel governance techniques with pre-existing ‘hard law’ requirements.

The UK government’s approach to regulation can be understood in the light of these factors. It emphasises that regulation should put less of a burden on industry and should not represent a block to innovation. Also, regulation should function at a higher, more strategic, level – implementing the dictum ‘less is more’.

The QCA’s approach to regulation reflects government priorities. It applies the following five principles of regulation:

- Proportionality (interventions are related to risk)
- Accountability (the public has a right to see what QCA does)
- Consistency (in judgements made; in data requested; in criteria used)
- Targeting (measures taken related to purpose)



- Transparency (open and visible)

The *Principles for e-regulation* can be understood in the light of this background. In implementing the principles, the regulators aim to:

- ensure that e-assessment strategy and operations are recognised as being robust
- guide operations, developments and innovative practice in e-assessment in a consistent way through principles of regulation
- support the extension of access to e-assessment opportunities for the benefit of learners
- identify and address parameters for success and areas at risk for innovative e-assessment strategy
- ensure that all regulation allows for flexibility, promotes and guides innovative development, and maintains the integrity, reliability and validity of e-assessment systems.

Thus, the *principles* are designed to maintain public confidence in e-assessment, whilst simultaneously supporting Awarding Bodies who wish to innovate and add value to qualifications through the use of technology.

The paper will give more detail on the scope of the regulatory principles, justifying why certain topics were covered but others omitted.

Next, some initial, exploratory research into the regulation of e-assessment will be described. This work will be based on a diverse range of data, including opinions gathered through questionnaire surveys, focus groups and similar approaches. Summaries of the main strands of opinion evidence will be given. Also, initial work to establish baselines to objectively illustrate the extent of uptake of e-assessment will be reported.

Finally, initial thoughts into the implications of e-regulatory research for the future of e-assessment more generally will be given. For example, comments will be made on issues such as:

- To what extent can early predictions about the benefits of e-assessment be justified?
- What can be done to ensure the successful and wide-scale implementation of e-assessment?
- How can risks that result from the use of e-assessment be minimised?



**THE MARRIAGE OF FREIRE AND  
BLOOM: AN ASSESSMENT  
PROTOTYPE FOR PEDAGOGY OF  
THE OPPRESSED AND HIGHER  
ORDER THINKING**

**Esyin Chew and Norah Jones**



# **The Marriage of Freire and Bloom: An Assessment Prototype for Pedagogy of the Oppressed and Higher Order Thinking**

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

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

## **Abstract**

The proposal delineates the problem of CAA and Bloom's taxonomy, summarising the pedagogical issues addressed by Freire and Bloom, and their relationships. The methods of data collection are explained concisely. The paper explicates several design elements of a system prototype, namely the *Learning HOTwatch v.1.0* based on the selected responses. The analysis and discussion makes its design meeting criteria such as reflection and substantive self-actualisation for high order level thinking. The preliminary architecture designed for the prototype is depicted and the similarity computation of case-based reasoning is suggested to use for the assessment computation. This proposal will be extended to provide further details in the short paper to be submitted.

## **Introduction**

There are various Computer-Assisted Assessment (CAA) applications in the market aimed to compliment the assessment process and to provide help for educators. The potential focuses are for the convenience of educators as well as the immediate feedback to the students. However, this results in a continuing problem: Does the question produced by such CAA application assess the learners at a higher order level?

Educationalists have been long aware of Bloom taxonomy (1956) which consists of six stages of cognitive thinking level. Bloom et al. (1956) found that most of the assessment questions require learners to think only at the lower level, which is information comprehension and memorising. Regardless the advancement of the innovation and intelligent in CAA, Higher Order Thinking (HOT) by Bloom et al. (1956) is, above all, a problematic reality in CAA. However, higher order thinking is a person's private experience, to which no

one else has direct access. The exam questions or assessment system may play a role in stimulating the higher order thinking skills for learner.

Thirty five years before Bloom, Paulo Freire with his famous publication *Pedagogy of the Oppressed* (Freire, 1980) critiques that the educator is the depositor who makes deposits whereas the students are the depository and they meekly receive, memorise and repeat (Connolly, 1980). The communication is a kind of monologue by the educator, people are taught to accept what is handed down to them by educator. Their understanding of particular knowledge is constrained to what they are told and then they just repeat what they are told during the exams. In such culture, learner are shaped to be silent and in ignorance (Bee, 1980). The learners are not given the opportunity to assess what has been assessed.

Conversely, Freire asserts that the aim of good pedagogy is to enable people to increase their understanding of their own objective conditions. Such understanding will inevitably lead the learner to assess the world as they climb out of the oppression in which they have been constrained (Barnard, 1980). He also captured the education qualities of what is to be human, and so education as a practice of freedom will remain pivotal for the realisation of the individual (Glass, 2001). Thus the learning process and angle is much wider and profound. Dialogue, reflection and communication to encompass this praxis are required (Connolly, 1980), and the role of the educator is to create such praxis, from theory to practical and also from lectures to reflections.

This perception is inevitably aligned with Bloom's Taxonomy (Bloom et al, 1956). The thinking level on knowledge, comprehension and application are more towards the conventional depository instruction method and lower thinking level whereas analysis, synthesis and evaluation are readily aligned to dialogue, reflection and assessment of the knowledge.

Likewise, Freire writes,

*'...acquiring literacy does not involve memorising sentences, words and syllables - lifeless object unconnected to an existential universe - but rather an attitude of creation and re-creation, a self transformation producing a stance of intervention in one's context.'* (Bee, 1980, p.42)

Hence, the aim of this paper is to study Bloom's and Freire's pedagogical praxis and to design an assessment prototype to embed such pedagogical issues into learning process.

## **Research Method**

There have been CAA applications research and design which are based on Bloom's taxonomy (King & Duke-Williams, 2001; Sitthiworachart & Joy, 2004; Paterson, 2002; Joy, Muzykantskii, Rawles, & Evans, 2002). Their research mainly focuses on how to assist educators in embedding HOT in question

design using CAA and to provide a set of exam questions with better HOT elements.

This research is an attempt to blend the educational theories from Bloom and Freire and it focuses on assisting the learners in an active and initiative manner.

This study incorporated the case studies with qualitative-quantitative interactive continuum methodology (Newman & Benz, 1998) due to its integrative and co-existent strengths of both qualitative and quantitative strategies. First, the arguments by Freire and Bloom are studied. In order to obtain the praxis in higher education institutions, three universities were visited and observed (one more to be visited in March 2007). Academic staffs and students from varying disciplines were interviewed and surveyed. The qualitative as well as quantitative data has been collected from their teaching and learning experiences.

The principal criterion in the selection of exemplary higher educational institutions was less “which HEI represent the totality but rather, “which group of HEI can gain better understanding for the research questions?” and ““which group of HEI reflect strong, both positive and constructive examples of the research interest?”. Given these criterions, a diverse group of HEIs and faculties were needed. For instance the traditional old universities and the new universities upgraded from polytechnic institutes, and the contrasting nature of disciplines related to technology such as Faculty of Computer Science and Faculty of Education; or the Faculty of Information and Communication technology and the Faculty of Humanities and Social Sciences are proposed for the criterion stated above.

To maximize the findings in a case study, a range of formal and informal data collection instruments are incorporated as listed below:

- Online and offline survey
- Recorded Face-to-face interviews
- Cases’ sites visits with direct observation
- Offline/ Online documentation, website, systems and data observations

The responses have been analysed and then act as an input to the design of a prototype which applies Freire’s and Bloom’s perception, namely, Learning HOTwatch v1.0.

## **Discussion, Analysis and Preliminary Design Issues**

The assessment of a learner on Bloom’s taxonomy is not only reflected in examinations, it can be assessed from the reflection of course work, tutorial, lecture, examination and the whole learning process. There are contrasting views offered from academics discussed next:

***Interviewee 1: Course work is the weak option in assessment because students can copy and whatever, and at the end of the day, the final exam is the true reflection. And it's always being driven like that...as long you have the assessment then you have the confidence that you actually truly assess the individual knowledge.***

***Interviewee 3: ...we are so much exam-oriented...because of this, teachers going into the class, what they think are, I want to cover the syllabus...I want to finish it and I want to give them exam and I want to drill my students until I got the model answers. Even during exam you must try to use that exactly word...to that extends for certain subjects...teachers maybe thinking assessment is always like we are teaching the students, and then we are assessing them, we give them test and exam at the end of the semester or the end of the term or at the end of the year... assessment actually can be done continuously...to assess our students in the process of teaching and learning and not assess them towards the end of the semester.***

In the conventional assessment method, the final examination is inevitably the way of imposing learners into HOT level. Freire further argues that pedagogy of the oppressed involves reflection and communication (Connolly, 1980). Such reflection process is a private experience and the process of learning is independent, no one else can assist and is not necessary carrying out only through conventional examination. This precisely stated by the following interviewee:

***Interviewee 7: It's not easy to teach the students the learning skills, the learning to learn by themselves. It depends a lot's on the students' ability to reflect on what happens.... to pick up the skill you have to do a lot's of reflection on your own.***

Thus, the key element of the Learning HOTwatch prototype is to provide the learner a continuous room for reflection by themselves and such assessment is not constraint to final examination but possibly the lecture, the course work and etc in the entire learning process. It provides a clear framework for learners to assess their own learning outcomes in Bloom's taxonomy boundary. With this framework in place, learners and educators are guided objectively and are able to assess the teaching and learning on an innovative manner. The insight gained by both learners and educators through this prototype may exceed what is generally available through traditional CAA-HOT assessment methods.

To demonstrate the learning reflection, general and simple externalization is substantive. The medium of externalization is not constraint to exam or lecture. It can be in any way:

***Interviewee 2: From my experience, I realised that when students express themselves, they are actually expressing what they have internalised. If I am giving a class, it doesn't matter what method I use, be a lecture or hands on or whatever method I use, what I do is normally...I force them to express themselves; it can be in any way. It can be in drama, it can be in song, it can be in poem, or just power point presentation, posters, modeling whatever..... They have to express themselves so that I can see what they have internalised. If they are not given a chance to express, to externalise what they have internalised, I would not know whether they have learnt. That's my technique, I make them externalise what they have internalised.***



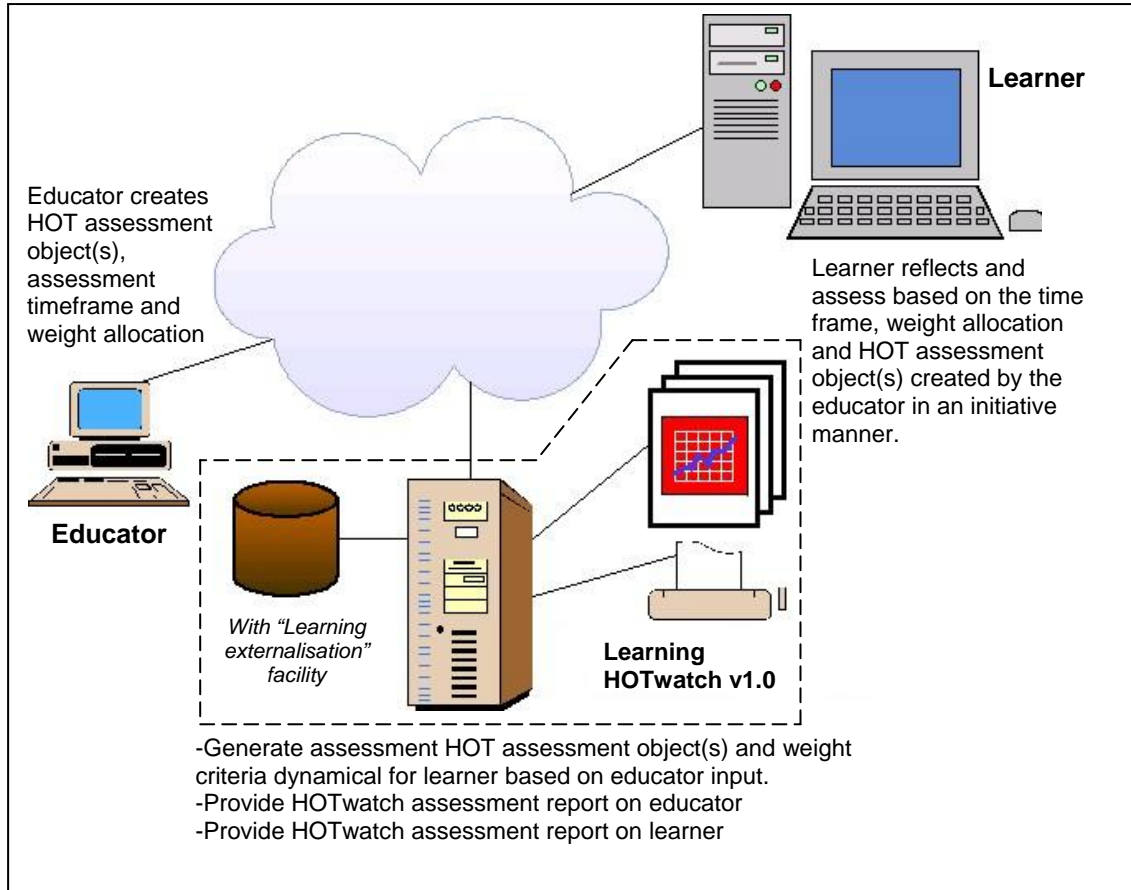
The Learning HOTwatch prototype aimed to achieve this by designing “Learning externalisation” facility for the learners to externalise what have been learnt. From the educators’ perspective, the role is changed from depositor and depository to facilitator and reflector. This prototype is not a system for setting up higher order thinking exam questions. It is a simple and general assessment tool to develop the learners’ contemplation. The learners may not have possibility to assess what have been assessed in a traditional CAA. It would be helpful if there is a system which allows the learner to express and to reflect their assessment of learning in higher order thinking rather than merely assessing their thinking skills. This complies the view from one interviewee:

***Interviewee 8: I want something like when people use your system, they will follow certain educational method and they will realise at last this is the learning process. In the class, when we ask students to google something and they will stuck when there are few thousand results return. There are some students who will choose the right website but some students will select the inappropriate site. Why is that so? Can we have one system to help those students who can't make a good choice to improve and know how to make a good decision? So, this system is educating the students to learn and not just information delivery.***

Paul (1993) suggests a model for the national assessment of higher order thinking to the United States Department of Education, Office of Educational Research and Improvement of the National Centre for Education Statistics. He claims that in addition to the assessment of learners’ skills in Bloom taxonomy, the model should be able to improve the instruction and enable educators to see what kinds of skills are basic for the future.

In such context, the Learning HOTwatch prototype is designed to concentrate on the ability leading to the improvement of instruction in a long run. At the same time it can be employed with maximum flexibility, in a wide variety of subjects and educational levels.

A preliminary model for the Learning HOTwatch prototype is depicted as the following:



**Figure 1.0: The Preliminary Design for Learning HOTwatch v1.0**

The algorithm of learning HOTwatch makes use of case-based reasoning, one of the expert system reasoning techniques to compute the result and report. Case-based reasoning is an attempt to apply the Analogical Reasoning to a practical problem (Leake, 1996). It is a methodology to model human reasoning without using rules for problem solving but matching algorithm. In summary, the Learning HOTwatch prototype itself corresponds to an if-then-else rule and it can be formulated into a complex computation model which is introduced in the Equation 1.0 and 2.0.

$$SIM(A, B) = [sim_1(a_1, b_1), sim_2(a_2, b_2), \dots, sim_i(a_i, b_i)] * w$$

where w = weighting i = assessment no.; a = educator's assessment b = learner's assessment

**Equation 1.0: Learning HOTwatch Similarity**

$$SIM(A, B) = 1 / f \sum_{i=1}^f sim_i(a_i, b_i) * W_i$$

where f = full weighting, w = weighting; i= assessment no

## Equation 2.0: Learning HOTwatch Similarity Computation Summary

### Conclusion

Freire insists that liberating pedagogy consists of reflection, critical dialogues and the acts of cognition, not the transfer of information from the depositor (the teacher) whereas Bloom taxonomy suggests the higher order thinking level that consists of analysis, synthesis and evaluation, which are readily aligned to the dialogue and reflection of the teaching and learning process

Overall, such teaching cannot be imposed from the top but instead should be carried out in a reflection process, shared investigation and in a problem-raising situation between educator and learners (Bee, 1980). The learner shall act as a subject and always possess critical thinking and maintain the dialogue with the educator, instead of being a submissive object in the learning process. Thus, this research is to design an assessment prototype, named, the Learning HOTwatch v1.0 which based on the pedagogical issues raised by Freire and Bloom, as well as the experiences from the academics. It will provide a bottom-up assessment via the process of articulation and reflection of higher order cognition by combining the considerations of two pedagogical approaches.

This proposal is a work in progress research in which the design and application flow of the Learning HOTwatch prototype will be illustrated in the short paper to be submitted later. Generally, review and discussion through sharing of ideas in web-based mediated environments has been implemented to facilitate forms of higher order reasoning (Wegerif, 1997; Crooks, 1994). In addition to this, the analysis and design of the Learning HOTwatch prototype aim to help educators distinguish more closely what they teach and by implication what they are assessing.

## References

- BARNARD, C. (1980) Imperialism, Underdevelopment and Education. In MACKIE, R. (Ed.) *Literacy and Revolution the Pedagogy of Paulo Freire*. London: Pluto Press.
- BEE, B. (1980) The Politics of Literacy. In MACKIE, R. (Ed.) *Literacy and Revolution: the Pedagogy of Paulo Freire*. London: Pluto Press.
- BLOOM, B.S., ENGLEHART, M.D., FURST, E.J., HILL & W.H.,
- KRATHWOHL, D. R. (1956) *Taxonomy of educational objectives. The classification of educational goals, Handbook 1: Cognitive Domain*. New York: Longmans.
- CONNOLLY, R. (1980) Freire, Praxis and Education. IN MACKIE, R. (Ed.) *Literacy and Revolution: the Pedagogy of Paulo Freire*. London: Pluto Press.
- CROOK, C. (1999) Computers in the community of classrooms. In K.Littleton & P. Light (Eds.), *Learning with computers: Analysing productive interactions*, London: Routledge, 102-118.
- FREIRE, P. (1980) *Pedagogy of the Oppressed*, New York: Continuum Publishing Company.
- GLASS, R., D. (2001) On Paulo Freire's Philosophy of Praxis and the Foundations of Liberation Education. *Educational Researcher*, 30(2), 15-25
- JOY, M., MUZYKANTSKII, B., RAWLES, S. & EVANS, M. (2002) An infrastructure for web-based computer-assisted learning. *Journal on Educational Resources in Computing (JERIC)*, 2(4), 1-19.
- KING T. & DUKE-WILLIAMS, E. (2001) Using Computer-Aided Assessment to Test Higher Level Learning Outcomes, *Proceedings of 5th International Computer Assisted Assessment Conference*, Loughborough: University of Loughborough, 177-187.
- LEAK, D.B. (1996) *Case-based Reasoning- Experience Lessons and Future Directions*. Menlo Park: AAAI Press, 1-34.
- McLoughlin, C. & Luca, J. (2000) Cognitive engagement and higher order thinking through computer conferencing: We know why but do we know how? In A. Herrmann & M.M. Kulski (Eds), *Flexible Futures in Tertiary Teaching*. Proceedings of the 9th Annual Teaching Learning Forum, 2-4 February 2000. Perth: Curtin University of Technology. <http://lsn.curtin.edu.au/tlf/tlf2000/mcloughlin.html>

- NEWMAN, I. & BENZ, C.R. (1998) *Qualitative-Quantitative Research Research Methodology: Exploring the Interactive Continuum*, IL: Southern Illinois University Press.
- PATERSON, J.S. (2002) Linking on-line Assessment in Mathematics to Cognitive Skills CAA, Proceedings for 6th CAA Conference, Loughborough:  
[http://www.caaconference.com/pastConferences/2002/proceedings/paterson\\_j2.pdf](http://www.caaconference.com/pastConferences/2002/proceedings/paterson_j2.pdf)
- PAUL, R. (1993) *Critical Thinking: What Every Student Needs to Survive in A Rapidly Changing World*, Dillon Beach, CA: Foundation For Critical Thinking.
- SITTHIWORACHART, J. & JOY, M. (2004) The Evaluation of Students' Marking in Web-based Peer Assessment of Learning Computer Programming, *Proceedings of the International Conference on Computers in Education, (ICCE 2004)*, Melbourne, Australia, 1153-1163.
- WEGERIF, R., MERCER, N., LITTLETON, K. & DAWES, L. (1997) Research Note: The Talk Reasoning and Computers (TRAC) Project. *Journal of Computer Assisted Learning*, 13(1), 68-72.



# **USING CAT FOR 11-PLUS TESTING IN NORTHERN IRELAND:**

## **WHAT ARE THE ISSUES?**

**Dr Pamela Cowan**





# Using CAT For 11-Plus Testing In Northern Ireland: What are the Issues?

Dr Pamela Cowan  
School of Education  
Queen's University Belfast  
[p.cowan@qub.ac.uk](mailto:p.cowan@qub.ac.uk)

## Abstract

This paper discusses the current concerns surrounding the psychometric properties of the Transfer Test used in Northern Ireland to select pupils aged 10-11 years old for a grammar school education. It highlights the lack of validity and reliability in the current selection system and offers computerised adaptive testing as the viable alternative for academic selection which reduces inequities associated with coaching and meets the international standards of validity and reliability.

## Introduction

Since 1947 the majority of schools in Northern Ireland (NI) have been operating a two-tier system of selective secondary education, commonly referred to as secondary and grammar schools. Places in the grammar schools are awarded on the basis of a Transfer Test, also known as the '11-plus test', taken at two unique times in the P7 year (final year of primary schooling). These tests are supposed to measure one or more of 'ability', 'achievement' or the 'potential to benefit from a grammar school education'. Gardner and Cowan (2005) completed a detailed psychometric analysis of these tests and revealed that they were lacking in validity and reliability as defined by the American Educational Research Association's *Standards for Educational and Psychological Testing* (AERA, APA and NCME, 1999) and that only 18 marks out of a possible 150 marks spanned the top grade (Grade A which secures a grammar school place) and the bottom grade (Grade D, commonly known as a 'Fail'). This paper outlines the concerns raised with the current Transfer Test and offers an analysis of the variety of alternative selection mechanisms currently under review by the educational bodies in NI including the use of Computerised Adaptive Testing for primary school pupils.

## The problems with the current 11-plus test

The Transfer Test comprises items addressing the NI Curriculum requirements in mathematics, English and science with the overall proportion

of marks awarded being in the ratio 26: 26: 23 respectively for these subjects. Since the Test determines the next stage in a child's education, they are viewed as 'high stakes' by parents, teachers and pupils. Although three subject areas are assessed, only one final grade (A, B1, B2, C1, C2 or D) is awarded summarising the scores in all three areas across both Tests. The technical fidelity of the Transfer Test was investigated using the following research questions:

- Are the tests unidimensional and therefore capable of differentiating pupils on a single construct, ability?
- How do children perform in the test?
- Are the tests successful in grading children accurately?

A random stratified sample of 52 primary schools of various sizes and school management types was used resulting in 1288 Test 1 scripts, 1270 Test 2 scripts and 623 Supplementary papers being returned. The Supplementary paper is only used if a pupil is absent from either Test 1 or Test 2. Confirmatory Factor Analysis, CFA, was used to test the null hypothesis that the proposed one construct model fits the observed data. The results were deemed to be 'safe' as the sample size was in excess of 200 (Boomsman, 1987). Using Joreskog and Sorbom (1989) the following limits were defined:

- $\chi^2 / df < 2.0$
- RMR < 0.05
- AGFI > 0.8

The one construct model, namely that the 'Test' captured 'the pupils' ability to benefit from a grammar school education', was tested for:

- the whole sample
- boys only
- girls only

and this model failed on the  $\chi^2 / df < 2.0$  criterion in every case.

Comparable tests for these three categories (whole sample, boys only and girls only) were conducted with the 3-construct model, that the test measured pupil performance in 'Mathematics', 'English' and 'Science/Technology', and for all three categories the 'goodness of fit' criteria listed above were met.

Nonetheless the disattenuated correlation coefficient between each of the 3 constructs indicates high levels of correlation (>0.8) as shown in Table 1.

Test/subject area	Maths & English	Maths & Science/Technology	English & Science/Technology
Test 1	0.853	0.915	0.898
Test 2	0.872	0.870	0.899
Supplementary	0.816	0.837	0.940

**Table 1 Disattenuated correlation coefficients**

A possible explanation for these high correlations is the phenomenon known as the Positive Manifold Effect whereby pupils no longer view the subjects as separate entities but as one interconnected unit called the 'Test'. As Primary School children are being taught by the same teacher, with the same teaching style and equal emphasis on each of the subject areas, the pupils find it difficult to distinguish between these three subjects resulting in a blurring of the boundaries. Messick (1989) advocates not collapsing different constructs or domains into a single measure even if they are highly correlated. Consequently, the results of the CFA show that the Transfer Test does not measure a single construct and treatment of the test score as a single measure, combining scores in the three subject areas, is open to question. As a result there is no evidence that the scores from the Test can be used to infer 'ability' or 'the potential to benefit from a grammar school education'.

### *Test performance and grades*

Frequency analysis of the Test scores showed that over 65% of pupils answered over 70% of the test items correctly yet this score equates to a Grade D ('Fail'). Clearly the 'easiness' of the Test lulls the pupils into a false sense of security in which they feel they have done well. On average pupils are correctly answering 70% of English and maths items and 83% of science items. This aspect of the test design is not acceptable for 'high stakes assessment'.

In terms of the grade allocations, the top 25% of pupils get a grade A which should secure them a place in a grammar school, the next 5% of pupils' scores are each awarded grades B1, B2, C1 and C2. The remaining 55% of pupils are awarded a Grade D (generally viewed as a 'Fail'). Due to the perceived 'easiness' of the test and the clustering of scores above 70%, the actual number of correct items required for a grade A in this sample was 123 out of 150, grade B1 ranged from 119 to 122 inclusive, while B2 was 116 to 118, C1 was 112 to 115 and C2 was 106 to 111. Scores of 105 or below out of a possible 150 marks were awarded a Grade D. With only 18 marks separating the highest and lowest grades it is imperative that the test is measuring accurately and the reliability is high. The Standard Error of Measure (SEM) of the test provides an indication of the precision with which the observed (raw) score reflects the pupils' 'true' score. With a SEM of 4.75 for the combined Test 1 and Test 2 scores, the 95% confidence interval reveals that the pupils' true score may be between 9 and 10 marks above or below their actual scores. Due to the close proximity of the grade boundaries, the potential misclassification for the Transfer Test is up to 3 grades and could effect over 30% of pupils (Please, 1971).

To date, the Transfer Test is not underpinned by any published standards of practice or technical fidelity. If international standards for educational and psychological measurement (AERA, APA and NCME, 1999) were applied to this test, two standards would raise particular concern:

## Standard 1.2 (on validity)

*The test developer should set forth clearly how test scores are intended to be interpreted and used. The population(s) for which a test is appropriate should be clearly delimited, and the construct that the test is intending to assess should be clearly described.*

## Standard 2.1 (on reliability)

*For each total score, sub-score or combination of scores that is to be interpreted, estimates or relevant reliabilities and standard errors of measurement or test information functions should be reported.*

Given the 'high stakes' nature of the Transfer Test in NI, issues of technical fidelity should be addressed. It is clear from this research study that serious concerns exist around the grading system embedded in the current Transfer Test. At present, debate is raging on whether or not academic selection should be retained and if it is, how can these issues of technical fidelity be addressed for future young people in NI.

## **What are the solutions?**

If academic selection is removed then parental preference will prevail. Parents will be encouraged to choose a school which best suits the needs of their child. This informed choice will be dictated by the Pupil Profile which is to be completed for every pupil in the primary school. The teachers will be required to complete and award levels in Communication (language and literacy), in Using Mathematics (mathematics and numeracy) and in Using ICT for each pupil. To supplement this record of academic achievements, the teachers will also be commenting on the pupil's Thinking Skills and Personal Capabilities, The Arts, Personal Development and Mutual Understanding, Physical Education, The World Around Us (Science, Geography and History), Religious Education, other interests and strengths and any other comments.

Although a database of pre-defined comments and phrases has been created, teachers are concerned about the time needed to complete each profile and also the subjective nature of the comments. Parents have been consulted and raised issues about the purpose and role of the pupil profile and how it could take into account the child's development over time – as soon as the profile is completed it is effectively out of date as the child will have moved on in his or her educational development. Parents viewed the pupil profile as guiding their decision-making regarding the 'best school' for their child however many parents found it difficult to interpret the content of the pupil profile and so training is needed for the parents.

If academic selection is retained or if schools are free to use academic selection if they wish then the key issues requiring attention are the international standards of validity and reliability of any tool used to select pupils into a post-primary school. However additional criteria have also been uncovered such as the need to accommodate pupils of all social and ethnic

backgrounds, the need to remove the pressure and anxiety associated with the Transfer Test as expressed by pupils, parents and teachers (Sutherland, 2000 and Save the Children, 2001), to allow parents to monitor their child's progress over the final years of primary education and to minimise the impact of coaching to ensure equity for all to all.

Some of the options under consideration include the use of NFER standardised tests in English and mathematics which will be administered on a specific day to all pupils, that is examination days like the current Transfer Test arrangement. This will not address the issues of validity for socially disadvantaged pupils and a new set of test papers would have to be created each year to prevent coaching. The stress on pupils, parents and teachers would remain as all tests would have to take place on a specific day.

A second alternative under consideration is the 'test when ready' facility of an Instructional Database Management System (IDMS). Unlike the NFER tests, IDMS has a large existing databank of test items for English and mathematics so pupils can have multiple attempts at the test. This option would reduce the feelings of pressure experienced on the 'test day' however it will not address validity issues for the socially disadvantaged child who has less access to coaching.

The University of Durham have been involved in the creation and use of an InCAS system (Interactive Computerised Assessment system) for English and mathematics which could be used to supplement the qualitative information provided in the Pupil Profile and to verify teacher assigned levels. Although the test is computerised, the system is not classified as a computer adaptive test and the website warns against making high stakes decisions based on the outcome of the assessments. This system would be a viable option for formative assessment and would assist the pupils, parents and teachers in determining a child's progress over time.

The final option under consideration is the use of computerised adaptive testing (CAT) which meets the standards of technical fidelity and removes the option of coaching to the test thereby ensuring equity to all pupils. Using Item Response Theory the test can be tailored to each pupil and so feelings of pressure and anxiety will be minimised as difficult questions are not administered. Also the pupils obtain instant feedback and due to the reduced assessment time, pupils can take the test when they are ready and repeat it as often as they wish. It is advocated that the CAT option could be taken at regular intervals in the final three years of primary schooling so parents will also have information about their child's progress over time and the results can be used for formative assessment prior to the P7 year. Although primary schools do not have dedicated computer suites, all primary and post-primary schools in NI are connected to the C2K network which offers secure 24/7 access via the external portal.

## Discussion

The use of computerised adaptive testing (CAT) for assessing pupils' mathematical attainment against the levels of the NI Curriculum for end-of-Key Stage assessment has already been demonstrated (Cowan, 1997) however its use for province-wide assessment leading to **high stakes** decision-making has yet to be piloted. International research in the use of CAT has focused on its role for admissions to US graduate programmes (GMAT and GRE), and formative and summative assessment in UK Higher Education courses however there appears to be a deficit of research into the role of 'high stakes' CAT with school age children and in particular with primary school pupils. If UK examination bodies are aiming to include on-screen assessment for all new qualifications and at GCSE and A level "*by 2009, e-assessment should certainly be normal, if not the norm, for thousands of students each year*" (Boston, 2004) then perhaps educators should consider preparing these students at an early age for high stakes assessment of this nature.

Worldwide, the pupil's age may vary for transferring from the first school to a second school however all pupils have the right to be assessed with validity and reliability as they make the transition to the next stage in their education. The name 'Transfer Test' may be synonymous to the NI context, however all pupils undergo some form of assessment as they move from one school to another whether it is an entrance exam or a decision made by parents or informed by teachers. Consequently the concept of using CAT to provide all pupils with the opportunity to demonstrate their ability with a high level of technical fidelity and consequently to be considered for a school which best matches their needs rather than based on assumptions from other adults, seems to be a child's human right.

## Conclusions

In terms of addressing the concerns raised in the Gardner and Cowan (2005) paper, CAT appears to be the only option meeting the call for international standards of validity and reliability. At the same time the nature of the CAT process facilitates a 'test when ready' approach which minimises stress and pressure on pupils and teachers alike while also minimising the impact of coaching. The content domain for the items will not distort the primary curriculum as teaching to the test is almost impossible. Pupils with special educational needs can be accommodated via the use of voice-overs and teacher time is not wasted creating lengthy pupil profiles with limited use to parents and post-primary schools. Since the tests are delivered online there is no need for extensive in-service training and moderation, a simple explanation of the -3 to +3 range for the scores is all that is needed for teachers and parents to interpret the pupil's test scores. Areas of strengths and weaknesses will be evident from the tracking facility within the software and detailed feedback against categories of test items can be provided for formative purposes. This system could be used over the final few years of primary education and weighted scores could be calculated to summarise

pupils' achievement over time. By making it an integral part of the primary school assessment system, pupils from all social backgrounds would have the opportunity to gain a place in the grammar school as 'opting out' would not be an option! So what is stopping the introduction of this CAT system as a means of academic selection in NI? At present it is 'cultural obstacles' (Hambrick, 2002) and politicians!

## References

- AERA, APA, NCME (1999) *Standards for educational and psychological measurement*. Washington, DC: AERA with APA and NCME.
- Boomsman, A. (1987) The robustness of maximum likelihood estimation in structural equations models. In P Cuttance & R Ecobs (Eds) *Structural modelling by example: applications in educational, sociological and behavioural research* (New York: Cambridge University Press).
- Cowan, P. (1997) Using information technology to assess and report mathematical attainment against a scale of standards. Unpublished doctoral thesis.
- Gardner, J. and Cowan, P. (2005) The fallibility of high stakes '11-plus' testing in Northern Ireland. *Assessment in Education*, 12, 2, 145-165.
- Hambrick, K A (2002) Critical Issues in Online, Large-scale Assessment: An Exploratory Study to Identify the Issues. Capella University
- Joreskog, K G and Sorbom, D (1989) *Lisrel 7: a guide to the program and applications*. Chicago, IL: SPSS Inc.
- Messick, S (1989) Validity. In R L Linn (Ed) *Educational measurement*, New York: American Council on Education. 13-103.
- Please, N W (1971) Estimation of the proportion of examination candidates who are wrongly graded. *British Journal of Mathematical and Statistical Psychology*, 24, 2, 230-238.
- Save the Children (2001) *Thoughts on the 11-plus: a research report examining children's experiences of the Transfer Test*. Belfast: Save the Children Fund.
- Sutherland, A (2000) Interviews with groups of Year 8 pupils. In T Gallagher and A Smith (Eds) *The effects of the selective system of secondary education in Northern Ireland*. Bangor: Department of Education.



# **WHAT'S NEW IN E-ASSESSMENT? FROM COMPUTER-MARKING TO INNOVATIVE ITEM TYPES**

**Nicola Craig**



# **What's New in e-Assessment? From Computer-Marking to Innovative Item Types**

Nicola Craig  
Pearson VUE

## **Abstract**

This session will be a brief exploration of the potential for more interactive and innovative item types in e-Assessment. Topics for discussion will include; how new simulations are being used by the medical profession; electronic marking of essays; and human marking of long answer questions. Whether you are already using e-Assessment and are looking for ways to innovate, or you are thinking of moving to e-Assessment and you would like to see what is possible, this session should have something for everyone.

Computer-Based Testing (CBT) has, believe it or not, been around since 1979. The prevalence for CBT started in the US with early adopters realising its benefits, such as Microsoft and Cisco in the I.T. market. CBT has since spread across the globe to encompass wider markets such as financial services, medicine and education. Here in the UK, it has seen growth within many bodies with a regulatory, licensure or academic basis. They utilise CBT as one form of a variety of assessment types to qualify and accredit their candidates and this number is on the increase year-on-year. A recent UK survey of more than 100 professional bodies found that two thirds had moved to CBT during the past two years – and that 63% expect a 'significant' increase in e-Assessment over the next five years\*.

So what is e-Assessment? It is the use of computer technology to present, record and mark responses to a test. Anyone taking a Computer-Based Test parks their pen and paper at the door and picks up a mouse instead. That said CBT isn't just about transferring paper-based questions onto a computer; it's more about harnessing a new way of testing that provides instant results, detailed feedback and an increase in the variety of item types.

Multiple Choice Questions (MCQs) have long been the preferred and statistically stable option when using CBT, but in this age of electronic innovations organisations are pushing the boundaries of what is possible in e-Assessment. Outside of what may be considered standard item types, e.g. MCQs, multiple-response, ordered lists, drag and drop etc the testing and assessment sector is seeing a greater interest adopting more sophisticated

---

\* A study of the use of e-Assessment by Professional Bodies, © 2007 Pearson VUE Ltd., FreshMinds Ltd.

technologies to create item types such as simulations, video clips and 3-D modelling.

So how useful are these items types and how much more can you achieve from an exam utilising these item types? Could expensive and time consuming 'practical' exams perhaps become a thing of the past?

In addition to the new and innovative item types, increasingly companies are looking to move their essay-based items over to computer. The benefits of this are huge as more and more candidates are used to learning and working with a keyboard rather than pen-and-paper. The benefits for markers is that the handwriting barrier no longer poses a problem and making assumptions and judgements on what the candidate 'may' have written, disappears.

An even greater benefit for these items is the potential to utilise an electronic human marking system giving you the ability to track and monitor your markers in real-time. This can give a greater consistency across grading and a rapid response and correction facility when a marker is going 'off track'. This can save an enormous amount of time re-grading at the end of a paper based marking event. Even with all of these benefits though how will the marker perceive the tool? Is it something that aids their marking processes or hinders them?

Alternatively machine-marking of essays is an option slowly increasing in prevalence, although with some caution. How effectively can a machine mark compared to a human? How much effort is involved in training the machine to mark to your specific criteria and over how many essay titles?

This presentation will take a brief look at some of the exciting item types being used in live testing and discuss the potential benefits of the results and examinations in this form. It will ask questions of you and encourage you to consider whether your own testing programme requires new item types, or if MCQ or essay-based exams are adequate to cover your syllabus, and indeed, if your organisation is using essay-type items then would there be scope to introduce computer-based marking?

# **REVIEW IN COMPUTERIZED PEER- ASSESSMENT**

**WILL IT HAVE AN EFFECT ON  
STUDENT MARKING  
CONSISTENCY?**

**Phil Davies**



# **Review in Computerized Peer-Assessment. Will It Have an Effect on Student Marking Consistency?**

Phil Davies

Department of Computing & Mathematical Science  
Faculty of Advanced Technology, University of Glamorgan, South  
Wales email: pdavies@glam.ac.uk

## **Abstract**

This short paper details work in progress that identifies an extension to the CAP Peer Assessment System that permits students to review the marks and comments of essays they've marked, having been allowed to view the comments of others who have also marked these particular essays.

The development of a compensation process that takes into account high and low markers is also discussed and whether the introduction of this review stage negates the necessity for this compensation process in the overall peer-assessment process.

Also presented is a review of the system in automatically allocating a 'mark for marking' that relates directly to the quality of the marker's work in both supplying marks and comments that match the quality of the marked essay.

## **Background**

Over the past seven years the CAP (Computerized Assessment by Peers) has been used as a tool to support the peer-assessment of both essays and multimedia presentations. This tool over this period of time has evolved from a basic marking tool that replicates traditional peer-assessment (Davies, 2000), to include anonymous communications between marker and marked (Davies, 2003) and the inclusion of menu driven comments and weightings to take into account subjectivity of the marker and automatic creation of a mark for marking (Davies, 2005). Throughout the various stages of development of this system the importance of feedback and quality of comments (Davies, 2004 & 2006) has been emphasised as being of great value to the owner of the essay. The rewarding of students for performing the marking and commenting in a qualitative manner has become one that has necessitated the introduction of a compensation process that automatically adjusts the marker's marks prior to the production of a compensated peer mark that acts as the final grade for a particular essay. Students have commented in the past that they find they have two major concerns in performing the peer-marking process:

- a) that they maintain consistency throughout the peer-marking process
- b) they are able to perform the 'task' well compared with other students in the group.

Following a successful internal grant application made to the University of Glamorgan's Teaching and Learning Committee, the opportunity arose to further develop the functionality of the CAP system to permit the students to amend their marks and/or comments for a particular essay having been permitted to review their previous marking of the essay. During this process they were also permitted to view the comments of their peers who had also marked this particular essay.

The new functionality of the system was then included in a trial study undertaken with a postgraduate cohort on a module teaching E-Learning within the academic year 2006-7. This paper describes the assessment process undertaken by these students and highlights the effect that this new functionality has had upon the peer-marking process.

Statistics are presented that show the increased time scales required for this aspect of the peer-assessment and whether the introduction of this review stage has had any subsequent effect on the quality and consistency of the peer-marking prior to the owner of an essay viewing their grades.

Discussion is also included that highlights the difficulty in providing an automatic reward for the peer-marking process undertaken by the students that maps to their quality of grading and commenting.

### **Assessment Description**

As part of their coursework assessment within a module teaching E-Learning, a postgraduate cohort of 13 students were requested to produce an essay in the form of a fully referenced RTF document that explained how to develop 'a distance learning Powerpoint presentation to teach 10 year olds something of a technical nature (in this particular assessment they'd been previously introduced to the Golden Ratio Phi as an example) but they were advised that this aspect of the assignment should not be subject specific. This report was to be addressed at the level of their peers and it was suggested that it was to be a maximum of three pages plus references. It was also requested that the main source of referencing be off the web (however some books & journals were to be expected). The reason for this being that in the peer-marking timescale permitted it would be difficult for a marker to be able to find book and journal references but as the CAP system supports an embedded web browser it would be easy for them to judge the relevant research undertaken by the essay developer. The students were given two weeks to research, develop and submit this essay.

Having performed this aspect of the assignment they were then expected to peer-mark and -comment at least six of their peers' work making use of the CAP system. The comments bank and criteria they used to assess the essays



will be explained later in this paper. Prior to the students undertaking the peer-assessment aspect of this assignment they were asked to use the marking system to self-assess their own work. This is an aspect of assessment that students in the past have found to be extremely difficult. The mark generated by this self-assessment process is not necessarily of great importance with regard to the outcomes of this assignment, however by performing this aspect of assessment it has been reported in the past that it has provided a means of the students

- a) getting used to the computerized assessment system
- b) having a way of creating a standard for themselves that they can use throughout the peer-marking process

The students were then given a week to perform the peer-marking process making use of the CAP marking system (Figure 1).

Figure One

The screenshot displays the CAP marking system interface. At the top, there is a navigation bar with tabs for Readability, Aimed at level of audience, Personal Conclusions, Referencing, Research and Use of Web, Content, Explanations, Examples, Case Studies, Overall Report Quality, Introduction, Definitions, Report Presentation, and Structure. Below this, there is a section for 'Go to Web Address' and 'Web Address'. A 'PULL DOWN MENU ONLY... DON'T TYPE HERE' section contains a list of references. A 'FREE TEXT COMMENTS' section has a text area for 'overall I think...'. A 'MARKS ALLOCATED' table is shown on the left, and a 'MINIMUM OF 10 WEB REFERENCES' section is on the right. The main content area displays a report titled 'Investigate the development of Grid Computing, and assess its possible future impact within the commercial sector'. The report includes an 'Introduction and Definitions' section and a section titled 'The differences between Grid and Cluster'. A green box with the number '1' is visible in the bottom right corner.

MARKS ALLOCATED	
Research Shown and Referencing	14 /30
Explanations and Examples	12 /20
Readability (material presentation)	13 /20
Subjective opinions (including justifications) and future	8 /30
<b>D Class Mark</b>	<b>47 /100</b>

Having completed the peer-marking process the students were then given a week to make use of the new review functionality added to the CAP system (Figure 2) which permitted them to view the comments of their peers concerning essays that they had previously marked. This paper reports upon the effect that this new review aspect has had upon the peer-marking process.

Figure Two

Phil Davies ...CAP Permit Marker Reconsideration

Readability | Aimed at level of audience | Personal Conclusions | Referencing | Research and Use of Web | Content | Explanations | Examples | Case Studies | Overall Report Quality | Introduction | Definitions | Report Presentation | Str...

Positive | Negative | Aimed at correct level | Aimed at roughly the correct level | Aimed too high but with good explanations

Web Address

0

**Pull-down Menu Comments**

Aimed too high without explanations  
Not referenced well, difficult to follow  
Good knowledge shown in subject area  
I have learnt from reading your report  
Enough detail  
Enough detail

Delete Highlighted Menu Comment

**Overview Comments**

You considered many of the right aspects in designing learning material but you failed to back up any arguments with references in the main text. Check your vocabulary - e.g 'prospective' not 'perspective'. Work on improving your language to convey masters level skills of analysis and critical thinking. Not tying your ideas to the research is expected at Masters level, this has let you down

**MARKS ALLOCATED**

Research Shown	5	/40
Explanations	22	/30
Readability and Structure	12	/20
Aimed at Correct Level	5	/10
<b>D Class Mark</b>	<b>44</b>	<b>/100</b>

Submit Revised Mark and Comments ONLY IF CHANGE MADE

Get An Essay I've Marked

Get Another Markers Comments

**Developing a Distance Learning PowerPoint Presentation Teaching ten (10) Year Olds a Topic Area like Phi**

**Introduction**

Developing any distance learning material can be a demanding task, subject knowledge and knowing the target audience you're aiming to teach is very important. During this brief report, we will be covering the broad outline of how to develop a PowerPoint presentation, from the initial gathering process, to the end presentation which will be delivered to the target audience. Obviously the development of any deliverable teaching material must follow a set pathway of development, which we will be covering in this report step by step, starting with the content information.

Introduction aimed to report title Fairly Readable Aimed at correct level Conclusions are very weak Not referenced well, difficult to follow Some interesting research found Reasonable attempt made to explain points Overall a good attempt Good attempt at structuring the report

Introduction is fine, but only explains the assignment title. Subject content last paragraph is a little too long. Accessibility not taken into account.

Subsequently the students were permitted to view the marks and comments of their peers with regard to their own submitted essays. They were allowed to view the median derived peer mark for their essay not the compensated peer-mark that would represent the final grade they were to be awarded for their essay.

In addition to this grade for their essay they were allocated a mark for the consistency shown in the peer-marking process that they had performed.

On completion of the assignment they were provided with a questionnaire requesting them to comment on how they had found the overall assessment process.

## CAP Application – Setting the Weighted Comments Bank

Prior to the self- and peer-marking of the assignment, the students were requested to develop an appropriate bank of comments that they could use within the ten categories used within the CAP menu driven marking system. Prior to the assessment being undertaken the students were offered the opportunity of replacing some of these ten categories and also to suggest suitable marking criteria for this particular assignment. Through discussion it was decided to leave the commenting categories as in the past, namely:

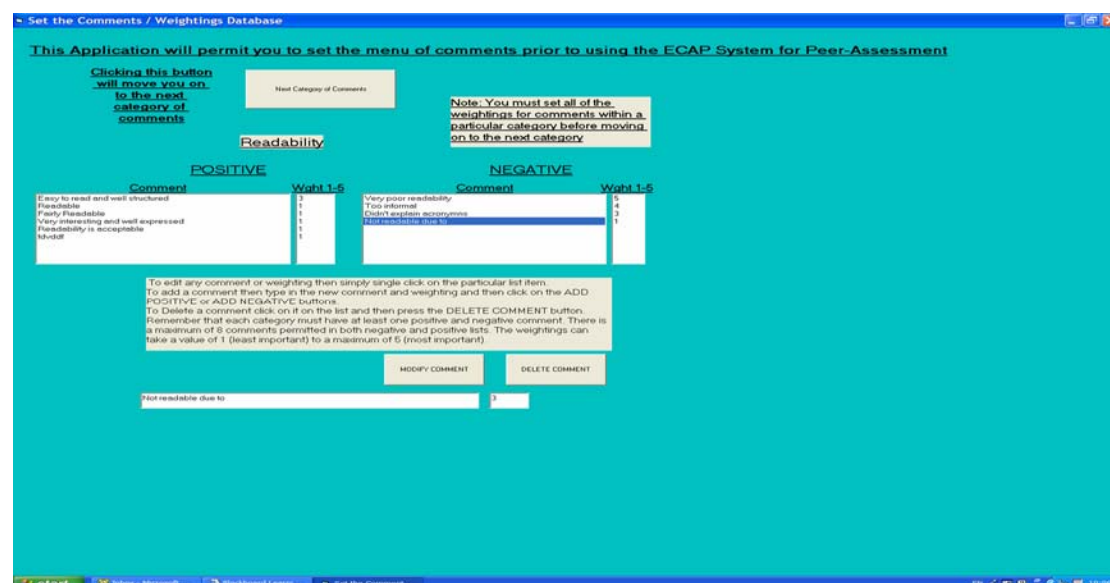
Readability, Aimed at correct level, Personal conclusions, Referencing, Research & use of web, Content & explanations, Examples & case studies, Overall report quality, Introduction & definitions and Report presentation & structure.

The marking criteria categories were:

Research Shown 40%, Explanations 30%, Readability & Structure 20% and Aimed at Correct Level 10%.

The students then made use of the Comments and Weightings setting application (Figure 3) to set comments that they felt suitable for their marking and including weightings per comment to include subjective importance for their commenting. This is described in more detail in Davies, 2005.

**Figure Three**



## Results

13 students undertook the assessment process however one of these did not complete the peer-marking process as requested. The result of this student has been included as the essay was peer-marked.

The overall compensated peer-mark generated for the essays was 60% with a standard deviation of 11.59. In order to generate this compensated average peer-mark for an essay, the possibility that a marking student is a 'hard' or 'easy' marker (often mapping to personal expectations) has to be taken into account. It would be unfair (unfortunate) from a student's perspective were they to be peer-marked by six hard markers compared to another student who was marked by six easy markers. In order to provide some form of compensation process, each marker has to be judged with regard to their average over- or under-marking methods. Each essay therefore needs to have a provisional average grade produced for it (the median is deemed to be a fairer reflection than the mean). Having created this, each marker's mark is compared against the average mark for the essay they've marked and an over- or under- average 'mark difference' is created. The essays now marked by this student are amended by this mark difference and a compensated peer mark is generated for each essay. Therefore the final peer-mark produced for

an essay is compensated taking into account the 'bias' shown by a marker. In the past uses of the CAP system the use of this compensation process has not really had a major influence upon the final grade produced, but certainly does allay the fears of students with regard to them being 'fairly' graded for their essays and not being disaffected by particular markers.

In past uses of the CAP system particular emphasis has been placed on the required quality of the comments produced mapping to the actual marks presented. Table 1 below shows the correlation between the compensated peer marks and the average feedback indexes for these essays (the quantified value taken into account the menu driven positive and negative comments):

**Table One**

+7	+6	+5	+4	+3	+2	+1	+0	-0	-1	-2	-3	-4
81	68	61	72		53		60	52		43		42
72		65						51				
								51				

As in past uses of the CAP system the above results on the whole indicate a very positive mapping of the comments received for an essay to the actual mark attained.

What this study has tried to ascertain is whether the students once offered an opportunity to review/modify their initial mark actually will do so. The preliminary analysis of this work indicates that out of a total number of 76 markings that took place there 41 're-markings' where either the menu driven comments and/or marks were changed. At this early stage of the analysis it is difficult to make any assertions as to in which way these re-markings have affected the overall peer-assessment results. It is possible that some students clicked on the 'submit a modified marking' button without actually performing a change?

The average time that a student took to mark an essay was 42 minutes (however this is not an exact timing that correlates to actual effort). It is interesting to note that the range of times included within this process was from 3-72 minutes. The students made good use of their menu comments with on average 16 comments being provided per marking.

Within the 41 're-markings' 26 of these actually resulted in a change of the original mark produced. The actual mark changes are detailed below:

+1, +9, +1, +2, +8, +6, +18 (71->89), +7, +6

-1, -2, -2, -8, -3, -5, -2, -1, -4, -6, -5, -7, -6, -3, -7, -7, -2, -5

Out of the 13 students involved in the study, 8 of these students made some form of amendments to their original markings, with 2 of these students actually 'modifying' all of their markings.

The conference presentation will provide a more detailed analysis of the findings of this study with regard to the re-marking process and whether by including this 'review' stage it has had any effect upon the actual final peer-marks that were produced.

A further aspect of this study is to attempt to automatically reward the students for the quality of their marking and commenting. A mark consistency figure has been generated that indicates the consistency shown by the marker (further explanation of how this is derived can be found in Davies, 2004).

It was decided in order to compare 'like with like', to map the consistency marks against the actual final compensated peer generated marks produced for the essays. In this way the 'range of abilities' of the students was used as a boundary to the percentage grade awarded.

For this group the average essay grade produced was 60% with a range of 81% to 42%. Thus the percentage points above the average being 21 and below being 18.

With regard to the mark consistencies produced the average being 4.87, with a range of 2.31 to 10.78 (keeping in mind that a low score is good whilst a high score is poor with regard to mark consistency). The resultant point range of a 'good' student below the average being 2.56 and that of a 'poor' student above the average being 5.91.

Therefore, mapping a good student's marking consistency to a good essay results in  $21/2.56 = 8.2\%$  for every mark consistency point below the average to be added to the essay average mark of 60%.

Similarly a poor student's marking consistency to a poor essay results in  $18/5.91 = 3.05\%$  for every mark consistency point above the average to be taken from the average essay mark of 60%

e.g. suppose a student has a mark consistency grade of 5.9. This is above the average mark consistency grade of 4.87, therefore it indicates a below average marking performance. To work out the percentage grade for this marking would result from an essay average (60%) – difference between the student's mark consistency (5.9) minus the average mark consistency figure (4.87) i.e. 1.03. This figure is then multiplied by the weighting for a poor result (i.e. 3.05%). Therefore the mark awarded to this student being  $60\% - (1.03 \times 3.05)\% = 60 - 3.14 = 56.86\%$ .

This method is obviously 'raw' and illustrates the difficulty in mapping an actual percentage grade to 'reward' the marking process in a qualitative manner.

Further analysis will take place and be reported upon concerning the feedback consistency and the effect that the re-marking has upon these consistencies.

From initial student comments this form of assessment has been met with general approval. A full analysis of the questionnaire results will be included at the presentation.

## **Conclusions**

At this early stage of the data analysis no major conclusions can be made as to the effect that the review stage has had upon the peer-marking process. The presentation will attempt to identify any significant trends, however these will be limited due to the small sample used within this study. Initially the results appear to indicate that the review stage does not have a major effect upon the peer-marks produced, thus the need for the compensation process remains.

At the onset of this study the author had mixed feelings concerning the possible outcomes of the introduction of the review stage. In past uses of the CAP marking system students have requested that they would have liked to have had the opportunity to re-assess their original markings, however the inclusion of this extra stage has been avoided in the past as it was felt that this may result in the students not setting their criteria for peer-marking clearly prior to performing marking due to the fact that they'd have a 'second chance'. The preliminary results appear to indicate that the students even though they knew that this second chance would be available took every care in their original marking (mainly due to the fact that they noted that they would be allocated a grade for performing this marking in a qualitative manner). The mark changes were relatively minor and appear to have little bearing on the overall results produced.

This addition to the functionality of the system has again met with the general approval of the students in that it has provided them with an opportunity to get a realistic appraisal of what their peer-assessment of the essays was in comparison with others within their group. Again it must be noted that this addition to the peer-marking system has resulted in an increase in the assessment time scales, and as such great care has to be taken in mapping an appropriate reward for the additional effort expected from the students in performing the peer-assessment process. This as in the past uses of the CAP system has to be mapped to the quality of the process not just the time taken.

## References

Davies, P. (2000), Computerized Peer-Assessment, *Innovations in Education and Teaching International (IETI)*, 37, 4, pp 346-355.

Davies, P. (2003), Closing the Communications Loop on the Computerized Peer Assessment of Essays, *Association of Learning Technology Journal (ALT-J)*, 11, 1, pp 41-54.

Davies, P. (2004), Don't Write Just Mark: The Validity of Assessing Student Ability via their computerized peer-marking of an essay rather than their Creation of an Essay, *Association of Learning Technology Journal (ALT-J)*, 12, 3, pp 263-279.

Davies, P. (2005), Weighting for Computerized Peer-Assessment to be Accepted, in Danson, M (Ed) *Proceedings of the 9<sup>th</sup> Annual International CAA Conference*, Loughborough University, pp 179-192, ISBN 0-9539572-4-1.

Davies, P. (2006), Peer-Assessment: Judging the quality of student work by the comments not the marks?, *Innovations in Education and Teaching International (IETI)*, 43, 1, pp 69-82.





# **BENEFITS AND OBSTACLES: FACTORS AFFECTING THE UPTAKE OF CAA IN UNDERGRADUATE COURSES**

**John Dermo**



# **Benefits and Obstacles: Factors Affecting the Uptake of CAA in Undergraduate Courses**

John Dermo  
Learning Technologist, University of Bradford, UK  
j.dermo@bradford.ac.uk

## **Abstract**

This short paper introduces and outlines a piece of research investigating the use of Computer Assisted Assessment (CAA) with undergraduate students, in order to identify the benefits of CAA as well the perceived obstacles to its adoption. It is hoped that ultimately this research will be able to inform the future use of CAA at undergraduate level, especially in blended learning environments. This research is currently in progress at the University of Bradford as part of the author's PhD and feeding into the university's Pathfinder project into e-assessment. The author hopes to be able to take advantage of the 11<sup>th</sup> International CAA conference to raise various issues related to this research project with his professional colleagues in order to receive feedback; this should enable decisions to be made on progress to date and inform how the research project may be developed in future.

## **Background and introduction**

The University of Bradford is striving to establish itself as a pioneer in CAA in the Higher Education Sector: the university has developed an exciting and forward-looking e-strategy and, as a Pathfinder Phase 1 institution, the University of Bradford will receive HEFCE funding under the HEA/JISC Pathfinder programme to develop e-learning for its maximum educational benefit, with a specific focus on embedding support processes for e-assessment with undergraduate students.

The National Student Survey has identified assessment methods and assessment feedback as important issues across the HE sector: at the University of Bradford these issues are now part of a debate which will lead to more comprehensive policy development regarding assessment. Based on a series of pilots, we believe that innovative e-assessment in general and computer-assisted assessment in particular can make an important contribution.

Developments in CAA at the University of Bradford so far include:

- Deciding on Questionmark Perception as our supported enterprise level software for online summative assessment

- Encouraging and supporting its use in formative assessment and feedback
- Centralising the administrative support for all summative assessments in our Examinations Office
- Implementing Questionmark Perception version 4.3 with a server configuration to ensure security and reliability

The investment in e-strategy will provide the support to expand physical facilities in this area; the focus in the Pathfinder project is developing the administrative and support systems. Building on small-scale pilots undertaken so far, the institution will develop the necessary systems to ensure reliable and secure large-scale implementation of CAA with first year undergraduate students so that we can subsequently roll this out to all students.

Whilst the University is encouraging staff to use its virtual learning environment (Blackboard) and Questionmark Perception to carry out formative as well as summative assessment, developments to date have been largely on an ad hoc basis, and with pioneering early adopters. It is recognised that a full-scale adoption of such e-assessment will require a combined commitment from the institution as a whole. This research should help to gather vital information from the key stakeholder groups to enable the institution to move forward in this area. It is also hoped that this research will be a useful contribution to the scholarship of e-Assessment uptake in Higher Education.

The focus of the research is primarily on high-stakes, summative assessment. Whilst much has been written in the literature about the use of CAA for formative purposes, relatively little research into summative e-assessment exists. The author feels that this is a challenging, interesting and important area, and is convinced that there will be considerable interest in the outcomes of this research in many HEIs across the UK.

The research does not restrict itself to objective forms of assessment, but also includes more open-ended subjective assessment and assignments delivered online. It hopes to cover innovative methods such as collaborative assessment, e-portfolios and even peer and self assessment, although it will be interesting to discover how these are perceived within the framework of summative assessment.

## **Methodology, design and methods**

This research project is descriptive and evaluative in nature, but hopes to inform subsequent more conclusive work. Of course, descriptive research is not simply the collection and presentation of facts and opinions, but it is the interpretation of the meaning or significance of what is described that is of primary importance. This approach is often criticised on the basis of the interpretation being affected by the researcher's own subjective opinion; it is therefore very important to have a carefully structured research design, with

clearly defined research questions as well as reporting results in clear and precise terms.

It is the firmly held conviction of the researcher that too much descriptive research is unsuccessful in its aims because researchers hurry into the data collection phase before they are sure that the research tools (e.g. questionnaires, interview questions) are ready for use. For this reason, the researcher is keen to spend extra time at the preparation phase to ensure that the data collected is valid, useful and reliable. It is hoped that this presentation at the CAA conference will be able to feed into this process.

The author is using primarily a qualitative approach to research. Educational research is not merely concerned with hard scientific facts and objective experimental hypothesis testing: the human factor in education can not be ignored, and attitudes and beliefs are of the utmost significance. Moreover, it is widely accepted that face validity is of fundamental importance in assessment, and e-assessment is certainly no different. The uptake of e-assessment will be greatly affected by the way in which students and instructors (as well as other key stakeholders) perceive the use of online assessment.

Given that we are interested in attitudes, opinions and beliefs, this is not an area that can be easily quantified. Also, the researcher favours a subjectivist, anti-positivist approach which suggests that educational issues cannot simply be described in objective, quantitative terms. This is reflected in the qualitative methodology favoured in this research. Of course, one of the challenges the researcher must face is how to reconcile a qualitative methodology with the need for generalisable results that are able to inform real word decision-making.

As for research design, this is a cross-sectional survey intended to capture an accurate description of stakeholder attitudes at a given point. It targets various groups of interested stakeholders in CAA: respondents are drawn from students and academic staff representing the full range of academic disciplines, administrators, invigilators, technical and learning support staff as well as management and financial and personnel departments and less obvious stakeholders, such as students' parents. In addition, the research is informed by external factors such as government policy, trends in HE and funding issues. This is a time when the institution is investing considerable resources in rolling out computer-assisted assessment as a fully supported service, so it is hoped that data gathered from this research will inform decision-making in the institution.

The initial phase of the research consists of focus groups and short interviews identifying key areas of interest. In conjunction with desktop research, this will form the basis of survey questions administered to all respondent groups. There will then be follow-up interviews in order to investigate key areas more thoroughly. The research does not set out to test a particular hypothesis, but is more descriptive in nature, intending to gather data to inform the decision-making process. The purpose of the initial focus groups and desktop

research is not to construct an a priori hypothesis, but rather to provide a focus for the research, to limit the scope of the data collection in a sound and reasoned manner. In this way, the descriptive survey can remain focused, and not simply gather data indiscriminately.

Another major design challenge the researcher must confront is how to gather stakeholder attitudes on issues which may be new to them: in other words, how do you find out what people think about computer-assisted assessment if they have never experienced such an assessment? It is anticipated that it will be necessary to include examples of computer-assisted assessment in action, so that the research subjects may be more informed in their responses.

### **Initial findings and looking ahead**

At present, the research is at the initial phase. The key areas of interest are being put to a full range of subjects so as to be able to inform the main survey questions to come, and at the same time the author is reviewing the literature. It is hoped that feedback from conference delegates will be able to feed into this process.

Research to date has identified the following as key issues to be explored further. The main drivers to have emerged so far include: savings in human and financial resources; improved reliability in marking; ease of production of results and item analysis data; ease of creation of different versions and randomised assessments; recycling assessments; positive backwash effect on teaching and learning; appeal to “digital native” students; possible benefits for recruitment and retention; potential of portfolio assessment; accessibility issues; encouraging good assessment practise concerning item banking and item analysis.

The obstacles emerging to date include: limited suitable task types; inability to assess higher level skills in a valid way; high risk of technical failure; initial outlay of time; steep learning curve for instructors; high cost of software licenses and support plans; difficulty in convincing examination boards and QAA concerning issues of quality; anonymous submission of assignments; security issues – e.g. passwords / collaboration / collusion / cheating / impersonation; item banking requiring more effort and time; technical expertise required of instructors; lack of immediate technical support; difficulties for administrators; difficulties for invigilators; training implications; accessibility issues; health and safety issues; difficulty of instructors in moving away from traditional task types; issues of task design; threat that CAA will be used to justify increased class sizes or staff reductions; lack of an agreed and enforced institutional policy; discrimination against “non-digital native” students; limited availability of Internet-connected computers at home, in halls of residences, on campus; availability of large computer rooms for examinations; lack of clear roles for technical services, administration, support services and departments.

It is immediately apparent that in these simple lists the number of obstacles is greater than the list of drivers, and many of these are already well described in the CAA literature. However, a key question to answer is whether the cumulative effect of the barriers outweighs that of the drivers. A key challenge facing this research is how to interpret the data in a meaningful way that can ascertain the degree to which e-assessment can add value to the learning experience. The research also needs to take into consideration the fact that some factors may work as drivers under some circumstances but as obstacles in others.

It seems that, whilst there is a lot of interest in CAA for formative assessment, many staff are still to be convinced of its value for summative assessment, and there is a great deal of concern about some of the perceived obstacles. However, it is to be noted that these are raw findings based on initial consultation with key stakeholders on the staff side. It should be very interesting to compare these findings with student data.

It is important to re-iterate at this stage that this research is based on an anti-positivist theory, and does not set out to test an objective hypothesis, but rather collect subjective data and set to make recommendations based on this. This will by necessity involve a certain amount of a posteriori theory construction: this will be one of the greatest challenges the researcher will have to face.

The researcher is keen to involve the input of other experienced practitioners and researchers in the field of Computer Assisted Assessment by means of this conference, and hopes to work this short paper up to a full paper submission for the next event in 2008.





# **THE USE OF INTERACTIVE ON-LINE FORMATIVE QUIZZES IN MATHEMATICS**

**Dr Judy Ekins**



# The Use of Interactive on-line Formative Quizzes in Mathematics

Dr Judy Ekins, j.m.ekins@open.ac.uk  
The Open University

## Abstract

In order to improve retention on Level 1 Open University mathematics, we are piloting short interactive internet quizzes. The OU package “Open Mark” is used, enabling students to receive instant feedback, where as previously they had to wait days or weeks. Students are allowed several attempts at each question, with appropriate teaching feedback after each attempt. At the end of each quiz, alongside the mark, relevant study advice is given to the student, including references to appropriate course material. Examples will be given.

Administrators can see all student attempts, helping in both modifying questions and feedback and for informing future initiatives. The quizzes are being evaluated using video of actual students “thinking aloud”, whilst attempting the quizzes.

User feedback on the pilot quizzes suggests that they are enjoyable as well as helpful to student learning.

Authoring and programming of quiz questions is time-consuming. However there is built-in variation, so that questions may appear in different guises for subsequent users and repeat attempts.

In the future, it is hoped to link the quiz feedback directly to pdf files of course materials and make these available together with the related quizzes on the OU’s “Open Content” web-site.

Keywords: e-assessment, mathematics, distance learning

## Introduction

The UK Open University (OU) provides supported distance learning undergraduate mathematics programmes. At level 1, there are two mathematics course modules: MU120 *Open Mathematics* and MST121 *Using Mathematics*. Several thousand adult students enrol annually on each. Student internet access has just become compulsory for administrative purposes and the University is adopting the MOODLE virtual learning environment. So we are keen to provide academic benefits for those who log on to the OU system.

MU120 is designed for students who have not studied mathematics for some time and/or who lack confidence. It introduces mathematical concepts in everyday contexts and it includes the topics of statistics, algebra, mathematical functions, regression, geometry, trigonometry, iteration, pre-calculus work, and mathematical modelling, together with using and programming a graphics calculator. Because its students come with a variety of previous mathematical skills, it has comprehensive preparatory materials. Students receive these materials when they register for the course, which may be several months before course start. Some students will need to spend a lot of time studying these materials, whilst others just take the allocated first two weeks of the course calendar to cover the material.

MST121 briefly recaps and continues many of the skills taught in MU120 and also introduces new topics, including sequences and series, conic sections, vectors, matrices, calculus and the computer algebra package MathCad.

Both MU120 and MST121 are studied over nine to ten months, with students submitting assignments approximately every four to six weeks. Of the students who start about 60% will complete. We are thus very keen to improve retention rates and keep as many as possible of the 40% non-completers.

### **The current assessment strategy**

Both MU120 and MST121 currently have a mixture of tutor-marked assignments (TMA), consisting of longish written questions, and computer-marked assignments (CMA), which are multiple choice tests. Most assignments are summative, i.e. the mark obtained contributes to the final overall mark. However both courses have a formative CMA on the preparatory work, which does not contribute to the overall mark.

Both TMAs and CMAs cover several weeks work. Each assignment has a cut-off-date, after which students receive comprehensive feedback on their work. However this may be a couple of weeks after they have completed the work and probably a month or more after they have studied the earlier topics covered in the assignment. Hence the feedback may not be as useful as if it were more immediate.

### **The usefulness of feedback on assessment**

The study of how assessment best supports learning is extensive. Gibbs and Simpson (2004) undertook a comprehensive review of the literature in this area and came up with 11 conditions for assessment to best support student learning. The current assessment strategy for MU120 and MST121 satisfies most of them, but falls short on one in particular:

The feedback is timely in that it is received by students while it still matters to them and in time for them to pay attention to further learning or receive further assistance. (Gibbs and Simpson, 2004, p. 172)

The pilot assessment is designed to rectify this. The medium chosen for the quizzes was the internet, in order to give speedy feedback at points where students would pay attention to it and use it in their learning. Brookhart (2001) discusses the differences between formative and summative assessment and Yorke (2001) discusses the role of formative assessment in retention in Higher Education. For the pilots, formative assessment was chosen to aid student retention. Buchanan (2000) emphasizes the role of feedback in fostering a meaningful interaction between student and the teaching materials, with particular emphasis on the use of web-based formative assessment. The OU's new web-based science assessment system "Open Mark", was adapted for the pilot mathematics quizzes, as it fosters such interactions.

### **E-assessment using "Open Mark"**

"Open Mark" is an on-line interactive assessment system, which has been developed at the Open University over a number of years, as outlined in Ross, Jordan and Butcher (2005). It aims to provide feedback to students, which is instantaneous, targeted and detailed.

Traditionally the OU has used multiple choice questions in CMAs, but "Open Mark" has broadened the range of question types. Thus enabling more skills to be assessed and making the assessment more interesting for students. Question types which enable plotting of points and lines on graphs, matching pairs, dragging and dropping words or symbols into appropriate places in mathematical expressions or text are available, as well as multiple choice and entering of numerical and algebraic answers. It is planned to integrate the "Open Mark" system into MOODLE, within the next year. It will then be Open Source.

"Open Mark" enables mathematical expressions to be entered easily and equivalent mathematical expressions are recognised as equally correct. Most questions can be designed in several variants that are randomly selected.

Students are allowed multiple attempts at each question (the maximum score diminishing with each attempt). They receive feedback after each attempt, tailored to the student's actual answer. The feedback after the final attempt usually includes a full worked solution or equivalent. We have also introduced a "hint" option, to help those, who don't know how to approach a question.

Examples of question feedback are: pointing out standard errors; telling the student if their answer is too large or too small; showing which parts of a multi-part answer are correct; and giving them hints. The feedback after the successive attempts often gives progressively more detailed hints. Details of the feedback mechanisms are given in Jordan, Butcher and Ross (2003), together with some of the technical aspects of the "Open Mark" system. There

is a demonstration web-site showing different types of question and feedback at <http://www.open.ac.uk/openmarkexamples>.

Upon completing each “Open Mark” assessment, students receive their marks and some appropriate study advice. References to the appropriate sections of the teaching materials are given, enabling them to quickly check on areas which need more attention.

A useful feature of “Open Mark” is the administrator’s reports, which show all responses for all users. This can be used on an individual level and on a macro-level to analyse responses, identify questions, where improvements might be needed.

### **The pilot mathematics “Open Mark” Quizzes**

For both courses, the principle is to provide short quizzes on coherent units of work. Students access the quizzes from their “Student home-page”. In order to explore the outcomes from different uses of “Open Mark”, the approach for MU120 and MST121 quizzes is different.

Each MU120 quiz has about six questions, based upon the one of the eight topics in the preparatory materials. The quizzes aim to help students assess their progress on a topic, as they complete it, at regular intervals, and to motivate them to continue with their studies. The quizzes use a variety of “Open Mark” question types, selected to best assess each skill. Students can attempt the quizzes as many times as they wish – the questions will be slightly different each time. Hence those who register well before the course start, will have plenty to keep them involved, where as those who register close to course start might attempt the quizzes just once in order to check their understanding.

The MST121 quizzes are designed to give the students practice in answering the type of questions on the summative CMAs. So questions are multiple choice. Students may tackle the quizzes throughout the course, after each chapter, and also use them in their revision for the final consolidation assignment

Current readers can try the MU120 quizzes themselves on the web-link: <http://mcs.open.ac.uk/mu120/>.

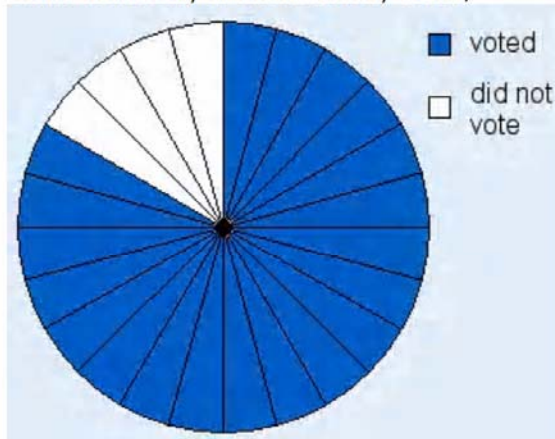
### **Examples of questions and feedback**

Here are several examples, including some of the feedback on incorrect and correct attempts.

## How's your Mathematics?

**Question 2** (of 7) • You have 3 attempts.

In a school class election for form captain, 24 students were entitled to vote. The figure below shows how many of them actually voted,



What is the fraction of eligible voters who actually voted in its simplest form?

fraction =  /

**Figure 1 Question on pie chart and simplifying fractions**

For the question in Figure 1, the hint gives:

“There are 20 shaded segments and so 20 students voted out of 24.”

Answers which are not equivalent to the correct fraction, receive a similar hint.

If the student gives the answer 20/24 they get the feedback:

“You have the correct number but the fraction can be simplified”.

Where possible all feedback is given to the right of the question, so as it is all on one screen, as in Figure 2..

### Question 6 (of 7)

The following temperatures were recorded by a meteorology station at 1am each morning in a particular week in March.

Sunday 3 °C	Monday 1 °C
Tuesday -1 °C	Wednesday 2 °C
Thursday -3 °C	Friday -1 °C
Saturday 0 °C	

Your answer is correct.

The temperature on Saturday was 0 °C.

Subtracting 3 °C gives  $(0 - 3)$  °C.

This is -3 °C which is the temperature on Thursday.

Which day is the day when it is 3 °C colder than on Saturday at 1 am?

Next

- ☐ Sunday ☐ Monday
- ☐ Tuesday ☐ Wednesday
- ☒ Thursday ☐ Friday
- ☐ Saturday

Check

Hint

**Figure 2 Feedback on a correct answer**

As in Figures 2, the feedback on correct responses always includes some working. However there is often some additional teaching in the feedback for correct responses, as well as for incorrect responses. For example a different preferred method may be given, as in Figure 3.

### Question 5 (of 7)



A shop has a half-price sale, which includes a wardrobe originally priced at £504. On the last day of the sale a notice appears on the wardrobe saying,

"Further reduction: 1/3 off sale price."

What is the final price of the wardrobe on the last day of the sale?

£ 168

Check

Hint

Your answer is correct.

Half the original price of £504 is £252.

1/3 of this is £84.

Subtracting £84 from £252 gives the final price of £168.

Alternatively you may have noticed that 1/3 off is the same as 2/3 left, so halve the original price, multiply by 2 and divide by 3 to get this result.

Next

**Figure 3 Alternative method, shown following a correct response**



The feedback on incorrect responses may be designed to make a student do some work, as in Figure 4 below, which shows a “drag and drop” question, in which the student has correctly dragged and dropped five definitions next to the relevant symbols, but has not attempted the other five.

Your answer is incorrect.

5 correct selections were made.  
Try again and correct those that are wrong or incomplete.

Note the following correct use of the symbols which you got wrong:

$$3 \leq 3$$

$$4 \geq 4$$

$$3^2 = 9$$

$$1/3 \approx 0.33333$$

$$3 \neq 4$$

Mathematical symbol	definition	
=	Is equal to	✓
√	Square root	✓
≤		✗
≥		✗
>	Is greater than	✓
<	Is less than	✓
^		✗
×	Multiply by	✓
≈		✗
≠		✗

Try again

**Figure 4 Feedback on an incorrect response**

Sometimes feedback on incorrect responses just gives a hint as to why the answer is wrong and reminds students of the technique, as in Figure 5 below.

**Question 2** (of 6)



A footballer earned £357720 one year.  
What is his pay correct to three significant figures?  
*Please enter your answer without commas.*

£ 357

Check

Hint

Your answer is incorrect.

You do not have enough digits in your answer. Maybe you need to add some zeros.

To round to 3 significant figures, look at the 4th significant figure (from the left). If it is 5 or more, round up on the first three digits, and zero from the fourth digit onwards. If it is less than 5, simply round down, i.e. zero from the fourth digit onwards.

Try again

**Figure 5 Question on rounding with feedback**

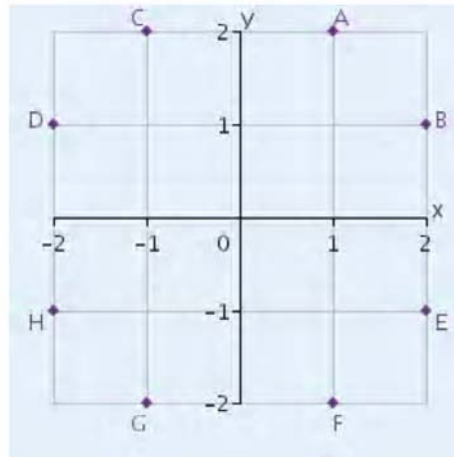
Students who get co-ordinates in the wrong order in the question below (Figure 6) will receive appropriate feedback.

**Question 4** (of 7)

On the right is a graph with a number of points marked on it.

What are the co-ordinates of the following points?

point B,	x: 1	y: 2	✗
point C,	x: -1	y: 2	✓
point E,	x: 2	y: -1	✓
point G,	x: -2	y: -1	✗



Check

Hint

Your answer is incorrect.

You have 2 out of 4 correct answers.

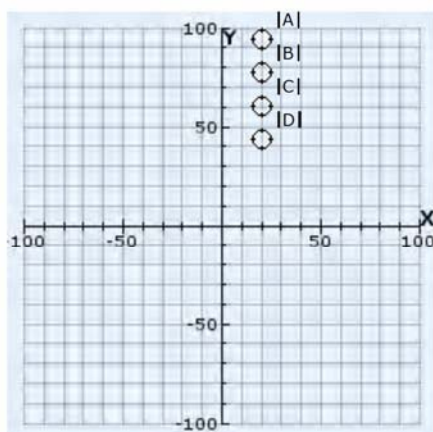
2 of your answers are correct but the wrong way round.

Try again

**Figure 6 Question and feedback on co-ordinates**

Students who get the question in Figure 7 wrong will be told which points are misplaced and reminded of the scale.

**Question 5** (of 7) • You have 3 attempts.



On the left is a graph with 4 points, A, B, C, D marked on it.

Using the mouse pick up each point and drag it to its allocated position given by the co-ordinates below.

- A (70, 20)
- B (70, -70)
- C (-70, 20)
- D (-70, -70)

Check

Hint

**Figure 7 A question on co-ordinates**

At the end of each quiz, students get a feedback page, which contains a summary of their performance, with an appropriate study comment, as in Figure 8, and a list of relevant references. The feedback page also gives students the opportunity to repeat the quiz with slightly different questions.

Here is a summary of how you did on this quiz:

#	Result
<b>How are your Powers?</b>	
1	Incorrect
2	Correct at 2nd attempt
3	Correct at 1st attempt
4	Incorrect
5	Correct at 2nd attempt
<b>Feedback</b>	
6	Correct at 1st attempt

### Overall score

7 (out of 15) 47%

You have some knowledge of the topics but you may wish to revise a number of areas. References are given below from MU120 Preparatory Resource Book A. This is available in pdf format on the course website.

Good luck with your studies.

### MU120 Preparatory Resource Book A references

#### Module 4

Question 1: Section 4.1.1 and 4.1.2

Question 2: Section 4.1.3

Question 3: Sections 4.2.1 and 4.2.2

Question 4: Sections 4.2.2 and 4.3.1

Question 5: Module 1 and module 4 overall

If you wish to re-run the diagnostic quiz with different questions, click the 'Restart entire test' button at the bottom of this page

Restart entire test

**Figure 8 Final summary and feedback page.**

### Feedback from Authors

The initial authoring of questions takes a similar amount of time to writing conventional multiple choice CMA questions. However because the feedback is more detailed and targeted, it takes longer to author. At the moment, the questions have to be programmed into the "Open Mark" system by somebody familiar with Java and so this is an additional resource, replacing publishing resource for print based assessment. Similar proof-reading is required for electronic and print, but because the feedback is more extensive, checking the interactive internet quizzes takes longer.

After the first user trials, the questions and their feedback were improved. This is an additional stage. However the finished product is much better than a conventional CMA. Another bonus is that, by including the variation facilities in “Open Mark”, one question authored is actually a set of similar questions, which lessens the need for further work in future years.

Individual students can be tracked on the administrative reports, which list all students who have attempted each quiz, their time on-line, all their responses and scores. The reports can also highlight problems. For example a summary of the question scores in Figure 9, highlights a problem with Question 4. On closer inspection and also from the “thinking aloud” video, it was found that the initial feedback on this question could be much improved.

#### Questions in test

The following counts include all who attempted a question, whether or not they finished the test.

#	ID	Taken by	Average score	Out of
1	<a href="#">mu120.module4.question01</a>	68	1.78	3
2	<a href="#">mu120.module4.question02</a>	61	2.02	3
3	<a href="#">mu120.module4.question03</a>	65	2.57	3
4	<a href="#">mu120.module4.question04</a>	60	0.72	3
5	<a href="#">mu120.module4.question05</a>	55	1.25	3
6	<a href="#">mu120.module4.question06</a>	49	3.00	0

**Figure 9 Part of an administrators report**

The reports can be used for analysing student errors as in Jordon (2006), but this is quite a lengthy project.

#### Feedback from Users

At the end of each MU120 quiz, there is a brief feedback question. In addition the quizzes are being evaluated using actual students “thinking aloud” as they complete the quizzes. The resulting videos are being analysed to see how the quiz questions and feedback stimulate their learning.

In some cases, opportunities for improving the quizzes were seen. For example, when a student missed out one of the two 60s in the calculation of the number of seconds in January, the feedback was not so appropriate, as this error was not anticipated.

From the administrator reports on the feedback questions and from the “thinking aloud” videos, it seems that users generally take between 5 to 30 minutes per quiz (less time on the earlier quizzes and more on the later ones). They generally like the immediate targeted feedback. However they are then critical, when the feedback is less specific to their answer. After the initial trials, it was sometimes possible to improve the feedback, but it is not always possible to anticipate every error.

The “thinking aloud” videos showed that the quizzes stimulate much learning. Students look up the relevant references, work on paper and use their calculators. They usually read the feedback carefully even if they had got the question correct. The students said that the quizzes stimulated their learning more than just doing exercises “from a book”, where the answers are in the back, or sending in their assignment answers and await feedback. If they got stuck or went wrong, they generally got more useful timely feedback from the quizzes, relevant to their actual answers. Students reported that they enjoyed the interactive quizzes, as well as finding them useful in checking their understanding and stimulating their learning.

## **Future work**

There is still much to be analysed in the use of the “Open Mark” mathematics quizzes. It is hoped to analyse the administrative reports further as well as the “thinking aloud” videos. An aim is to examine how different types of question and feedback stimulate learning, highlighting relevant aspects for future authors. The assessments themselves can be improved and the results of the project considered by course teams for new and rewritten course modules. In particular the rewrite of MU120 is about to commence.

After the end of this year’s presentations, the retention rates of MU120 and MST121 students using the quizzes will be compared with those who do not use them.

Once “Open Mark” is integrated into MOODLE, it is also hoped to make the MU120 preparatory materials, together with its set of quizzes, available on the OU’s Open Content initiative, for all to use. In particular people contemplating registering for the course will be able to study the preparatory material in their own time beforehand and receive helpful tailored feedback on the attempts at the quizzes.

## **Conclusion**

The “Open Mark” system has enabled us to pilot the use of interactive internet assessment with OU level 1 mathematics students, to make feedback more immediate and useful within student learning. Initial trials suggest that users find the quizzes fun as well as useful for their learning. Authoring of quizzes is more time-consuming initially than the traditional CMAs, but less work subsequently. Students generally liked the shorter quizzes with detailed tailored feedback.

There is still work to do in analysing the videos of students “thinking aloud” and the administrator reports on student responses, which can inform improving teaching materials.

It is hoped that the increased motivation and improved feedback will lead to better student retention but this can only be judged after course end next year.

The pilots have stimulated discussion of mathematics assessment and much of interest in the Faculty. Hopefully this will provide a stimulus for us to use the internet and the new MOODLE VLE to improve our assessment and teaching.

## References

Brookhart, S.M. (2001) Successful students' formative and summative uses of assessment information, *Assessment in Education*, 8 (2), 153-169.

Buchanan, T. (2000) The efficacy of a World-Wide Web mediated formative assessment, *Journal of Computer Assisted Learning*, 16, 193-200.

Gibbs, Graham and Simpson, Claire (2004), Does your assessment support your students' learning?, *Journal of Teaching and Learning in Higher Education* (on-line), 1, 3-31. Retrieved August 19, 2005, from: <http://www.glos.ac.uk/adu/clt/lathe/issue1/index.cfm>

Jordan S, Butcher P, & Ross S (2003) Mathematics Assessment at a Distance, Maths CAA Series: July 2003

Jordan S (2006) The mathematical misconceptions of adult distance-learning science students, CELT- MSOR Conference, Loughborough September 2006

Ross S, Jordan S, & Butcher P (2005) On-line instantaneous and targeted feedback for remote learners, in Bryan C & Clegg K V (eds) *Innovation in Assessment*, Routledge Falmer, London (2005)

Yorke, M (2001) Formative assessment and its relevance to retention, *Higher Education Research & Development*, Vol 20, No 2





# **MODERNISING ASSESSMENT: THE USE OF WEB 2.0 FOR FORMATIVE AND SUMMATIVE ASSESSMENT**

**Bobby Elliott**



# **Modernising Assessment: The Use of Web 2.0 for Formative and Summative Assessment**

Bobby Elliott  
SQA

## **Abstract**

This paper considers current assessment practice, looks at the impact of the Internet on today's learners, and explores ways of modernising assessment to narrow the real or perceived gap between the everyday lives of students and the assessment practices that we impose on them.

## **Assessment 1.0**

At its most basic level, assessment is the process of generating evidence of student learning and then making a judgment about that evidence. Current assessment practice provides evidence in the form of examination scripts, essays or other artefacts.

### *Characteristics of Assessment 1.0*

For the purposes of this paper, 'Assessment 1.0' can be thought of as assessment practice from the beginning of the 20<sup>th</sup> century until today. Throughout this period, assessment exhibited the following characteristics:

- mostly paper-based
- mostly classroom-based
- very formalised (in terms of administration)
- highly synchronised (in terms of time and place)
- highly controlled (in terms of contents and marking).

These characteristics were largely unchanged during this period; a school master from 1907 would feel at home in an examination hall in 2007.

This system of assessment has served us well. The highly centralised, top-down, command-and-control assessment system matched the kind of society that existed throughout most of the 20<sup>th</sup> century. Its stability has engendered widespread public confidence in the examination system in the UK (QCA 2006)<sup>i</sup> and maintained national qualifications as the primary means of employee selection and progression to Higher Education. The system is also widely understood by its users (students, parents, teachers, university

admissions staff, employers and politicians), being relatively unchanged from generation to generation.

### **Assessment 1.5**

A more up-to-date form of assessment has developed in the last ten years, which involves the use of computers in the assessment process. 'E-assessment' embraces 'e-testing' (a form of on-screen testing of knowledge) and 'e-portfolios' (a digital repository of assessment evidence normally used to assess practical skills).

#### *Problems with assessment 1.0 – and 1.5*

In recent years, traditional assessment has been the subject of criticism. The current system is struggling to cope with the demands being placed on it. It was designed to filter students by ability for the purpose of employment or university selection – not mass accreditation of student achievement. Because of its bureaucratic nature, it's expensive to run and doesn't scale well. Awarding bodies' costs are rising and these are being passed onto schools and colleges, which complain about the rising burden of examination fees. It's also inflexible, organised around examination "diets".

In addition to these practical considerations, there are educational and political concerns. Some educationalists claim that the current assessment system encourages surface learning and "teaching to the test". Instead of instilling genuine problem solving skills, it fosters memorisation. Examination papers that appear to pose "deep" questions are answered using rote memory – memories that are acquired by students under pressure from parents who are keen to see their children gain qualifications, and drilled by teachers who are seeking to meet targets. Employers complain that, in spite of rising achievement (DfES 2006)<sup>11</sup>, young people are not gaining the skills that are needed in the modern workplace – skills such as collaboration, team working, problem solving, adaptability and creativity. Teachers complain about the rising burden of time spent carrying-out and marking assessment– and which reduces the time available for "real learning". Students themselves complain that the only time that they are required to undertake extended writing is during an examination.

These criticisms are not confined to paper-based assessment. E-testing has been criticised for crudely imitating traditional assessment. Vendors of computer-based testing systems boast about their systems' faithful reproduction of the paper experience. These systems typically support a limited number of question types (almost always selected response questions) and, at best, crude simulations of traditional tasks. Most contemporary e-portfolio systems, likewise, set-out to mirror the existing curriculum, effectively little more than online storage for students' work, with a highly content-focussed (rather than student-centred) approach to assessment.

“In 21st century learning environments, decontextualised drop-in-from-the-sky assessments consisting of isolated tasks and performances will have zero validity as indices of educational attainment.” (Pellegrino, 1999)<sup>iii</sup>

These criticisms of e-assessment mirror the criticisms of VLEs – that they simply seek to copy conventional practice: the “primacy of pedagogy” as Cousin (2004)<sup>iv</sup> described VLEs’ slavish simulation of the traditional classroom rather than seeking to capitalise on the unique opportunities afforded by technology. Cousin observed that: “VLE environments (*sic*) tend to be skewed towards the simulation of the classroom, lecture hall, tutor’s office and the student common room.” Similarly, most contemporary e-assessment systems are skewed towards the simulation of the class test and the examination hall. Or, to paraphrase Cousin, they re-enforce the “tyranny of testing”.

Both paper-based and computer-based assessments are perceived by students as something external to them; something that is “done” to them; something over which they have no control. And the assessment instrument itself is considered contrived and artificial: just a hurdle to be jumped – not part of their learning. Assessment 1.0 (and 1.5) is also intensely individualistic. Assessment activities are done alone, competition is encouraged, and collaboration (or “cheating” as it is known in the world of Assessment 1.0) is prohibited.

Not ideal preparation for the ‘networked information economy’.

## Web 2.0

Meanwhile, the Internet is evolving. ‘Web 2.0’ is the name given to the current state of development. Anderson (2006)<sup>v</sup> describes “six big ideas behind Web 2.0”. These are:

- user-generated content
- the power of the crowd
- data on an epic scale
- architecture of participation
- network effects
- openness.

For the purposes of this paper, four of these ideas are of particular relevance.

**User-generated content** refers to the ease of creating content. Web services such as MySpace, Blogger and YouTube have made it easy to create content – and more and more young people are doing exactly that, with social networking sites becoming a significant part of contemporary culture.

The **power of the crowd** refers to the collective intelligence that can be harnessed from large groups of people. The basic premise is that, subject to

certain conditions, a large group of knowledgeable (but non-expert) users can make better decisions than any individual expert. Web services such as Digg and Wikipedia are cited as examples of this collective intelligence.

**Architecture of participation** is based on the twin ideas that Web services must be easy to use (thereby encouraging participation) and must be organised in such a way as to improve as more people use them. Google Search is a good example of both since it is very straight-forward to use and its search algorithms (which are proprietary) learn from the results of previous searches (although the precise means are not known). An aspect of ease-of-use is the idea that not only is new content easy to create but it should be easily created from pre-existing content or easily combined with the contents of other web services (“mash-ups”).

**Openness** not only refers to the use of open source software for many Web 2.0 services but also the philosophy of the free sharing of information and resources among users, making it relatively straight-forward to capture and share information or resources, such as embedding a YouTube video in a blog.

### *Digital natives*

It is in this environment that today's students are living and learning. In *Digital Natives, Digital Immigrants* Prensky (2003)<sup>vi</sup> argued that there was a fundamental distinction to be made between today's learners and those of the past due to “the arrival and rapid dissemination of digital technology... an event which changes things so fundamentally that there is absolutely no going back”. He labelled these new learners “digital natives” and contrasted them with “digital immigrants”: “The single biggest problem facing education today is that our digital immigrant instructors, who speak an outdated language (that of the pre-digital age), are struggling to teach a population that speaks an entirely new language”.

Today's learners are also known by other names. Diana Oblinger (2003)<sup>vii</sup>, of Microsoft, calls them the ‘Millennial generation’: “Millennials exhibit distinct learning styles. For example, their learning preferences tend toward teamwork, experiential activities, structure and the use of technology. Their strengths include multitasking, goal orientation, positive attitudes, and a collaborative style”. From the student's perspective, “Net Geners” are “academically driven... we refuse to accept elders' speeches or sermons at face value... our technological savvy makes us smarter, easily adaptable, and more likely to employ technology to solve problems” (Windham, 2005)<sup>viii</sup>.

### *Different learning styles*

A common set of characteristics emerges from the literature on the digital native with respect to their learning styles. These are:

- skilled use of tools
- active learning rather than passive receiving of knowledge
- authentic learning experiences rather than contrived tasks

- task (not process) oriented
- just in time learning
- search not memorise
- utilise social networks
- doesn't know answer but knows where to find it
- Google not libraries
- collaborate not compete.

When tasked with an assignment, a young person is likely to look-up Wikipedia, search for relevant information on Google, seek help from their friends via Hotmail or MySpace, finally pulling together the resulting information into a coherent document using a range of web-based and desktop applications. Unless, of course, the assignment is the same as last year's, in which case a simple e-mail to a friend (or someone else in their extended social network), requesting last year's answer, will be sufficient for these goal-oriented learners.

#### *Disjoin between classroom practice and real world behaviour*

The above scenario sidelines the formal teaching and reference material that the student is meant to use. There is a growing disconnection between the lives of students inside and outside of the classroom. "Schools should not expect students to leave the 21<sup>st</sup> century in the cloakroom; for example, many schools do not allow e-mail, instant messaging, mobile phones or blogging" (Owen *et al* 2006)<sup>ix</sup>. And the list of prohibited technologies is growing. Twist and Withers (2006) describe the ways in which young people really learn as the "hidden curriculum" – the "informal digital spaces", such as MySpace and MSN, which students routinely use for social and educational purposes.

## **Assessment 2.0**

This paper proposes an update to Assessment 1.0. The updated system will embrace the Internet and, more specifically, Web 2.0 – particularly the four "big ideas" described above. It seeks to bring the 21<sup>st</sup> century into the examination room.

#### *Characteristics of Assessment 2.0*

The type of assessment activity best suited to the digital native would exhibit some or all of the following characteristics.

- **Authentic:** involving real-world knowledge and skills.
- **Personalised:** tailored to the knowledge, skills and interests of each student.
- **Negotiated:** agreed between the learner and the teacher.
- **Problem oriented:** original tasks requiring genuine problem solving skills.
- **Socially constructed:** using the student's social networks.

- **Collaboratively produced:** produced in partnership with fellow students.
- **Recognise existing skills:** willing to accredit the student's existing work.

And the type of evidence that best fits this type of assessment would be:

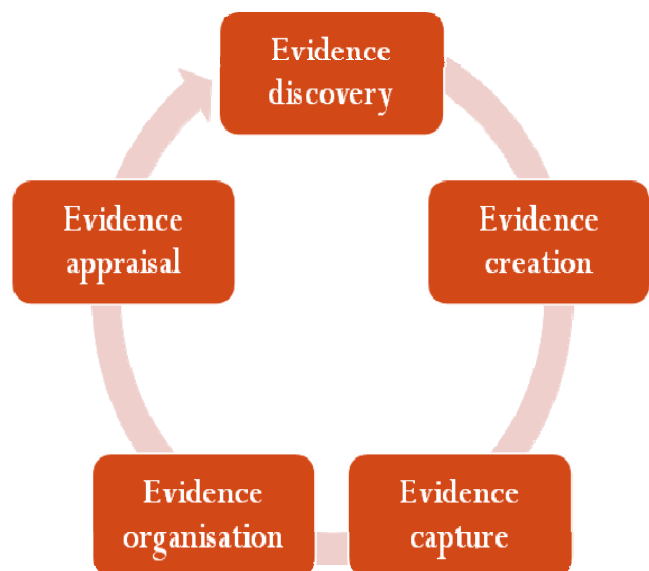
- **naturally occurring:** already in existence or generated out of personal interest
- **digital:** such as e-mail, instant message logs, blog posts, wiki contributions, audio and video recordings
- **multimedia:** existing in text, audio and video format
- **distributed:** may be scattered across various sources (such as web sites, blogs, inbox, iPod).

For example, an Assessment 2.0 task relating to language skills would permit the student to explore a topic of personal interest to them, negotiating the precise parameters of the task with their teacher, working in conjunction with fellow students, and recognising the student's previous writing on the subject (such as their MySpace page). The evidence could be in a number of digital formats such as e-mail conversations, IM logs, blog, web site or wiki.

#### *How Web 2.0 can be used for assessment*

Assessment is about evidence generation. The diagram below illustrates how evidence is traditionally produced.

Evidence has to be discovered (when it already exists) or created (when it does not). The resulting information has to be captured and organised. And, once it is coherent, the evidence has to be assessed. It is straight-forward to relate this model to Web 2.0. The following table illustrates how a range of Web 2.0 services can be used for one or more of these stages. For example, a contemporary web-based e-mail system (such as Google Mail) can be used as a repository of every e-mail message you ever send or receive – which could be an Aladdin's Cave of assessment evidence.





The following table relates a number of Web 2.0 services to the assessment cycle.

Web service	Example	Cycle	Use
Personal portal	Netvibes	Evidence organisation	Combining items on single page
E-mail	Google Mail	Evidence storage	Searching e-mail archive for evidence
Blog	Wordpress	Evidence organisation	Recording activities
RSS	Bloglines	Evidence discovery	Subscribing to evidence sources
Social bookmarking	Del.icio.us	Evidence capture	Capturing URLs
VOIP	Skype	Evidence capture	Talking and chatting
Wiki	Wikispaces	Evidence creation	Collaborative writing
Instant messaging	MSN	Evidence discovery	Chatting
Search engine	Live Search	Evidence discovery	Locating information
Online storage	Box.net	Evidence organisation	Saving and storing information
Data capture	Clipmarks	Evidence capture	Selecting and storing information

Downes (2006)<sup>x</sup> describes the combination of Web 2.0 services for learning as “personal learning environments” (PLEs), arguing that the PLE is a “recognition that one-size-fits-all approach of LMS [VLE] will not be sufficient to meet the varied needs of students”. Assessment 2.0 posits Web 2.0 as a **personal assessment environment** in recognition that the one-size-fits-all approach of e-assessment systems will not be sufficient to meet the varied needs (and interests) of candidates.

### Advantages and Disadvantages of Web 2.0 for Assessment

Given that Web 2.0 is Life 1.0 for most students, it is an easy fit for most young people. They are already using Web 2.0 services as part of their everyday lives. Recognising their MySpace page or their YouTube video or their Odeo podcast seems only “fair” to them. And in doing so, it would reduce the perceived chasm between education and “real life”. It would also provide an incentive to learners; instead of artificial tasks involving “ancient” practices (such as hand-writing or using the library), assessment could provide real challenges using real tools – the same tools that they will use in the workplace. Web 2.0 is inherently collaborative and the antithesis of Assessment 1.0’s obsession with individuality – and collaboration is a skill much sought after by employers. Web 2.0 services are also inexpensive (or free), easy to maintain (since it is maintained by someone else), and very scaleable (in fact, the more users the better). The alternatives (dedicated e-

testing systems and e-portfolios) are expensive, difficult to maintain, (usually) proprietary, and quickly become out-of-date.

There are drawbacks. Older students (our digital immigrants) aren't using Web 2.0 services – or, at least, not routinely. They don't have MySpace pages or YouTube videos to be plundered for accreditation of prior learning. And they may lack some key Web 2.0 skills (such as search skills) and attitudes (such as a willingness to share). Assessment 2.0 also poses challenges for teachers – who are often the epitome of the digital immigrant. Not only might they lack the IT skills needed to understand Web 2.0 services but they may lack the knowledge and experience required to appraise students' work produced using these tools. They also lack the rubrics required to assess Web 2.0 skills, such as collaboration and team work. Group work is notoriously difficult to assess – so difficult that most awarding bodies prohibit it from high stakes assessment. Yet, it is at the core of Web 2.0 and a crucial skill for the workplace. Authentication is another challenge for awarding bodies in the world of Assessment 2.0, with the myriad sources of digital evidence and collaborative inputs making it a challenge to authenticate an individual piece of work.

## **The Future**

It's impossible to confidently predict the future. But there are certain themes that emerge when you review the international literature relating to the future of education and technology. With regard to education, there is a consensus about the following:

- greater focus on education as a key differentiator between countries in the global economy
- growth in learning at all stages in your life (the “forty year degree programme”)
- the emergence of new skills to better fit the networked information economy
- greater role for e-learning (and particularly mobile learning)
- move towards personalised learning (and, by corollary, personalised assessment)
- greater recognition of informal learning.

In tandem with these educational developments, the next decade may see the emergence of ubiquitous computing and Web 2.0 will evolve into Web 3.0. “Ubiquitous computing” describes a state of pervasive computing where digital devices are embedded into everyday life to such an extent that we are unaware of their existence. And Web 3.0 will consolidate the “big ideas” of Web 2.0.

“Educational institutions may be reconfigured from monolithic institutions to resources operating across different domains (e.g. home, school and community); educational practices may prioritise collaboration and reflection

rather than the acquisition of knowledge; and educational goals may be re-imagined as personal and bespoke rather than mass-industrial and one-size-fits-all. At the heart of these visions are personalisation, collaboration and learning to learn.” (Owen et al 2006)<sup>9</sup>

If you combine these developments, you see a digitally rich environment, where learning will take place in multiple locations (at school, at home, on the bus), at a time to suit the learner; where learning is personalised – in fact, a world where the distinction between learning and living is blurred and assessment evidence occurs naturally as part of the student’s everyday personal and professional endeavours.

## **Conclusion**

Assessment is often accused of preventing educational change. The critics accuse high stakes assessment of dictating the educational system and stifling innovation. So, if education is to change, that change has to be led by the assessment system.

One of the ways assessment can evolve is to embrace some of the characteristics of ‘Assessment 2.0’. That means embracing Web 2.0 and the digital environments that students inhabit. Doing so would present a challenge to teachers and awarding bodies. Teachers would have to up-skill to understand Web 2.0. Awarding bodies would have to face the challenge of creating rubrics for assessing difficult to measure skills, such as collaboration, and confront issues such as plagiarism. Both teachers and awarding bodies would have to embrace digital evidence in all of its forms and set more authentic tasks that genuinely challenge (and engage) students.

“It will not be easy but the next generation will create new models of scholarly publishing and learning regardless of whether we choose to participate. The only question will be what role we carve out for ourselves.” (Thompson 2006)<sup>xi</sup>

We’re talking evolution – not revolution. There is a place for Assessment 1.0, 1.5 and 2.0. We just need more of the latest version.

## References

---

QCA. GCSEs and A level: the experiences of teachers, students, parents and the general public.

DfES. Johnson Welcomes Rising Achievement. Retrieved 22 April 2007. [http://www.dfes.gov.uk/pns/DisplayPN.cgi?pn\\_id=2006\\_0119](http://www.dfes.gov.uk/pns/DisplayPN.cgi?pn_id=2006_0119).

Pellegrino, JW (1999). The Evolution of Educational Assessment. *William Angoff Memorial Lecture Series* (ETS).

Cousin, G (2003). *Learning from Cyberspace*. JISC.

Anderson, P (2007). What is Web 2.0? *JISC Technology and Standards Watch*. pp. 14-26.

Prensky, M (2001). Digital Natives, Digital Immigrants. *On the Horizon* (NCB University Press, Vol. 9 No. 5, October 2001).

Oblinger, D (2003). The Millenials. *Educause*. August 2003.

Windham, C (2005). The Student's Perspective: Educating the Net Generation. Chapter 5. *Educause E-book*.

Owen, Grant, Sayers and Facer (2006). Social Software and Social Learning. *Futurelab, Opening Education Series*.

Downes, S (2007). Learning Networks in Practice. *Emerging Technologies for Learning*, Volume 2, Chapter 2 (2007). Becta.

Thompson, J (2006). Is Education 1.0 Ready for Web 2.0 Students? Nova Southeastern University

# **CAN TERTIARY E-ASSESSMENT CHANGE SECONDARY SCHOOL CULTURES?**

**Andrew E. Fluck**



# **Can Tertiary eAssessment Change Secondary School Cultures?**

Andrew E. Fluck  
Faculty of Education  
University of Tasmania  
Andrew.Fluck@utas.edu.au

## **Abstract**

Tertiary eAssessment has a crucial role to play in secondary schools. This paper reports on a pilot project which replaced written examination papers by CD-ROMs. Whilst the traditional supervised fixed-time assessment process was preserved, students were able to use modern digital technology to create text and graphical (drawing) responses, without collusion. The paper suggests that such innovations at the tertiary level of education may eliminate barriers to transformation of schooling through ICT at the secondary level.

## **Introduction**

"If the exam is on paper, then that's how we'll teach!" Australian secondary schools are trapped between policy pressures to use computers in classroom practice, and the reality that students progress into universities where hand-written examinations are crucial. This paper addresses a barrier to change in schools by demonstrating how university examinations can be transferred onto computers. This transition makes sense for tertiary students, since much of their learning is conducted using online materials blended with face-to-face activities (Mogey, 2006). Winkley & Osborne (2006) have written about the distributed development of item banks and reticulated examination setting, but also note:

"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)."

Such an approach addresses misalignment between learning and assessment technologies (Ashton & Thomas, 2006) and is therefore more likely to gain acceptance. Thus it was adopted for the pilot project described below.

## Literature

### *Tertiary assessment as a barrier to ICT in schools*

“There is cautious ground for optimism [about ICT in schools]. 83% of teachers interviewed in schools said that they believed that ICT can raise standards. *Yet, we wonder, why is this a belief instead of a reality after the investment of so much in terms of both money, time, commitment and energy in ICT over the past twenty years?*” (Reynolds et al., 2003)

There is a clear problem in schools where policy drivers promote the use of ICTs but the reality of classrooms mitigates its impact. If Australian schools are to aspire to the transformative uses of computers, they need assurance this will not impede pupils as they pass into tertiary studies. The transformational role of ICT in schooling has been underlined by the four types of use identified for the Australian government (Downes *et al.*, 2002, p.23):

- Type A: encouraging the acquisition of ICT skills as an end themselves
- Type B: using ICTs to enhance students' abilities within the existing curriculum;
- Type C: introducing ICTs as an integral component of broader curricular reforms that are changing not only how learning occurs but what is learned;
- Type D: introducing ICTs as an integral component of the reforms that alter the organisation and structure of schooling itself.

Pervasive high-stakes hand-written examinations in the tertiary sector are a major disincentive for changing current text production methods in schools.

If ICT use in schools is restricted to Types A or B, then the full benefit of high technology investment will be limited. However, for uses corresponding to Types C or D, school cultures will need to change markedly. This transformational view of ICT in schools requires a rethink about curriculum content, the applicability of previously established learning outcomes and criteria to the future lives of student, and even the structure of schooling itself (Fluck, 2003; Fluck, 2005; Tinker, 2000).

This project hypothesised that transformational thinking in schools is inhibited by a range of factors from teacher skill with ICT, infrastructural capacity, cultural inertia, perceived equity and curriculum constraints (Enerson, 1997). At other levels, such as awarding bodies, return on investment uncertainties and candidate authenticity are among the barriers to acceptance (Chapman, 2006). One critical inhibitor was considered to be school sector attitudes to formal assessment processes, particularly those associated with pre-tertiary qualifications and beyond to undergraduate examinations. This was given credence by the operations manager of the Tasmanian Qualifications Authority who related discussions with 'laptop' schools, in which all pupils have a personal computer. These schools have not lobbied for pre-tertiary



entrance examinations to be undertaken using computers, because they know students need experience in the hand-written testing they later encounter in University assessments. This can be seen as a crucial attitudinal obstruction for the adoption of ICT-dependent information handling in schools.

By providing tools to eliminate this inhibiting factor in a small range of cases, there will be an opportunity to study the potential of ICT in schools that are subsequently able to change their cultural approaches. This has been demonstrated by a recent trial in Victoria, Australia which allowed Year 12 (age 17-18) students to complete English examinations using computers (Maslen, 2004). The author's personal conversation with those involved suggests that students who were previously regarded as having illegible handwriting - a very large proportion of adolescent males - did exceedingly well in the keyboard environment.

### *Online or Offline?*

As a lecturer in a pre-service teaching course, I have noted the displacement of supervised examinations by unsupervised home assignments, to the point where less than 10% of student grades in some degree courses derive from rigorously identity authenticated assessment. Students receive mixed messages from school and university assessment, where much is done in collaborative team settings, and these are not easily distinguished from work required to be completed individually. Students assigned a personal online quiz have been observed gathering a team of friends to assist in the completion of the assessment. The online nature of the process appears to blur the line between collaborative and individual assessment, since the examiner has little control of the context in which it is undertaken.

This emphasis on online learning is often accompanied by the requirement for students to submit assignments which have been printed rather than handwritten. Lecturing staff are gradually moving into a multi-media mode, and some are asking for the submission of assignments on CD-ROM or to a content management system. This digitalisation of tertiary learning has not been matched by internal examinations. Sometimes these examinations are considered 'high stakes', as they represent up to 70% of the total mark for a unit being studied. The University of Tasmania has a well orchestrated examinations system: exam halls are booked; papers are securely drafted, checked and printed; furniture is moved into place; exam periods are invigilated by employed supervisors; papers are securely distributed for marking; and results collated for posting. Very little technology is allowed into the examination hall. Mobile phones are banned, calculators are required to be identified on the exam paper, and a few dictionaries may be permitted. One of the consequences is that local physiotherapists do extremely good business during the examinations. It is quite evident that students are put under increased stress because of the nature of the handwriting process required, which is so very much removed from the rest of our teaching and learning practice.

### *Strategic importance of eAssessment*

ICT and associated skills are seen as strategic, and nationally important for economic, pedagogical and social reasons (Hawkrige, 1989). The Australian Information and Communications Technology in Education Committee AICTEC (2006) has recognised the strategic importance of eAssessment in its detailed report on an identity management framework, noting “the effective management of e-Education including the areas of eLearning, eAssessment and eReporting between providers, learners and parents (in the case of learners in the compulsory school years). The importance of this issue at a national level has been increasingly recognised in the work undertaken by a range of national groupings and initiatives - e.g. AICTEC, the Education & Training Statistics Advisory Group and the Student Mobility Working Group.” Similar sentiments have been expressed by the UK Department for Education and Skills [DfES] (McGill, 2006).

When it comes down to it, the value of a certificate from a learning institution is worthless if it can be obtained by an individual who has completed less than the entirety of the accredited programme. The pathway taken in this pilot project has been to suggest an intermediate solution for eAssessment enabling proctored examinations to be taken using CD-ROMs instead of printed examination papers.

Why go for an offline solution? There are two reasons. The first is that current proprietary web-based testing solutions such as Exambient (Blackboard, 2007) or Software Secure Secureexam (SecureExam, 2007) require responses to be formed within the browser context or using producer-defined applications. Therefore the choice of software candidates can access is severely limited to those tools provided by the testing environment web-page. This is not acceptable if students are to really demonstrate their expertise using a wide range of popular software tools such as Audacity (audio editing), The GIMP (image manipulation), Mathematica (mathematical analysis), FreeMind (concept mapping) and other highly complex applications. The second reason for choosing a CD-ROM based solution was the flexibility it gives examiners for making the decision about connectivity. One paper may use a CD-ROM stripped of all internet connectivity software functions, forcing candidates to use the facilities of the isolated workstation. The afternoon paper on the same computer may allow access to a selection of five critical web-sites. By configuring these options when the master CD is created, the decision is left to the examiner.

This is not a solution to the difficult problem of online identity management in certification situations (Fluck, 2005b; Pescaru & Holotescu, 2002), but only a step towards it. The system described in this paper is for proctored or supervised examinations, where the identity of candidates is verified by reference to documentation upon entry into the examination room, and where inter-candidate communication is strictly monitored and generally forbidden. A more general solution for examinations outside this on-site context using may emerge from this proposal, and a combination of approaches involving biometrics (UK Passport Office, 2004), third-party proctoring and a controlled

IT environment might be subsequently developed for completely on-line testing.

## **Study**

Most computers have at their core an operating system and applications stored on a non-volatile medium such as a hard disk drive. When the computer is switched on, this set of instructions is loaded into RAM and subsequently controls the machine's behaviour. An alternative is to provide a complete operating system on CD-ROM, a medium which cannot be altered during the examination. At least three such systems are available: Lindows, Knoppix (Knopper, 2004) and Ubuntu. They can be used on almost any modern computer (Mac or PC) just by inserting the appropriate disk and switching on the computer. When everyone is using identical copies of the same CD-ROM, the result should be equitable. These systems (others may be available or could be created) are open-source (Fluck, 2004a; Office of Government Commerce, 2004): therefore students can legally be given copies to take away and practice with, facilitating their familiarity.

Over the past two years a small pilot project at the University of Tasmania has provided proof-of-concept. In brief, we have been able to assess a cohort of 167 students using an on-computer examination system. Students were issued free copies of the examination system to boot their own or a Faculty computer from (without the paper!) a month beforehand to enable them to become familiar with the environment. This allowed the students to spend considerable time practicing with the examination environment since the CD-ROM was built from open-source components. This built personal confidence and ironed out some initial problems well before the examination day.

On the day of the test, computers in a conventional laboratory were started from copies of the same CD-ROM containing a 'live' operating system and the examination file. The 'live' operating system CD-ROM provided us with an environment which was pre-engineered to suit the circumstances of the examination. On such a CD-ROM we can prepare an operating system which has no network functionality, no tools for inspecting the local hard disk drive and no other software except that strictly required for the examination. The environment we selected included Open Office (which is similar to Microsoft Office), and a program called 'GIMPshop' which compares with Adobe PhotoShop for image manipulation. Since these applications use open file standards, the response files produced will be accessible over a longer time span than alternatives using proprietary formats, as required by the National Archives of Australia (Zymaris, 2004, p. 26). The preparation of this CD-ROM is analogous to the printing of an examination paper.

The questions were fairly unremarkable, being almost exactly the same as one would expect on a conventional examination paper at this level. The topic was classroom pedagogies using ICT. Figure 1 illustrates some typical questions. They were a mixture of short answer, image manipulation, and attitudinal types. Only a few were of the knowledge-based multiple choice

variety. This flexibility to provide an examination close to the paper-based original, but encouraging the use of sophisticated software tools was hoped to produce a test of high-level thinking whilst retaining the digital environment to which students were accustomed.

**Question 7:**

What are the dangers of a 'Cut & Paste' culture?

**Question 10:**

How do robots provide an example of 'Problem solving' with ICT?

**Question 22:**

Describe one way to create an animation, and one reason for using this technique in a teaching setting.

**Question 26:**

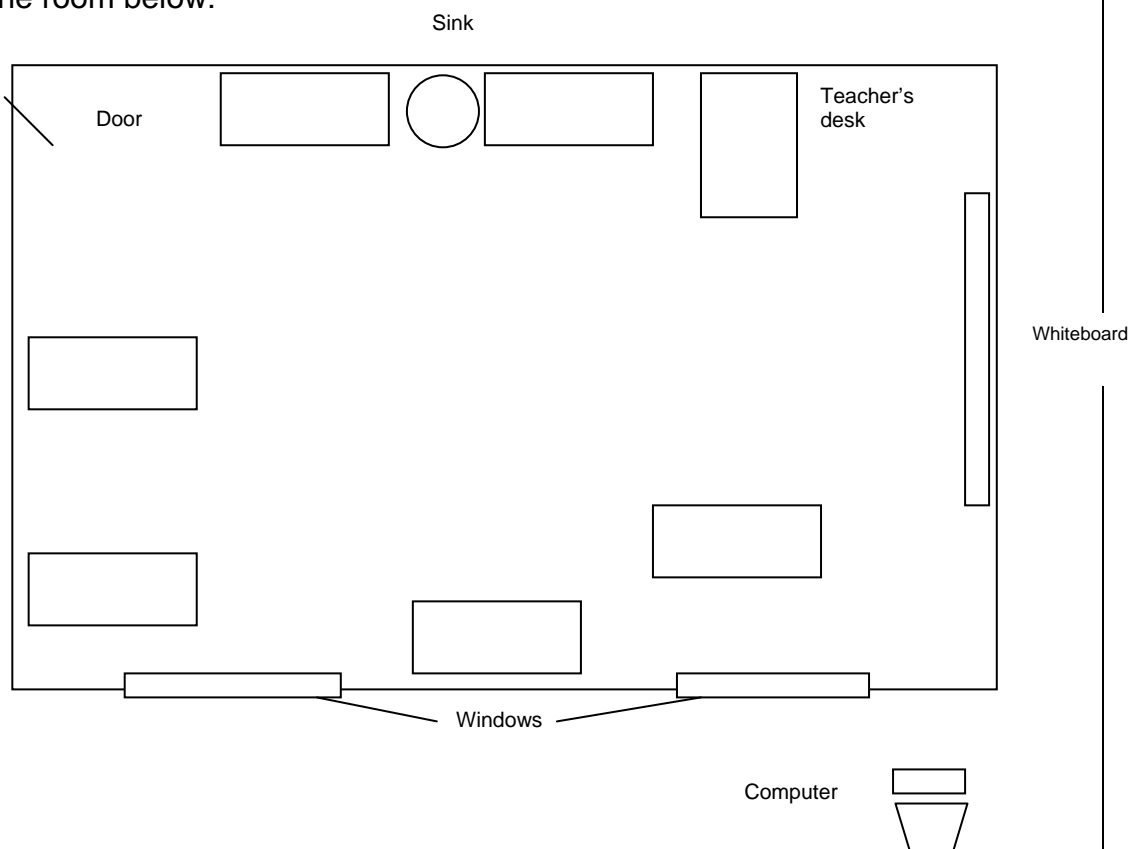
How do you suggest a student with no arm control take national literacy tests?

**Question 33:**

Concept-mapping software can be used to visually plan a unit using a theme-based approach. Provide a diagram which shows the start of such a plan.

**Question 32:**

If you had 4 classroom computers, show where they would best be situated in the room below:



**Figure 1: Sample questions from pilot examination**

The inclusion of a unique artistic feature on the desktop background for each examination helped non-technical supervisors ensure the correct operating environment was present on each candidate's computer. In the pilot, the computer file of the examination question paper was loaded onto the desktop using a USB data-stick, but in future years we will pre-burn this onto the 'live' CD-ROM. In a similar way, we collected completed scripts using another USB data-stick. They were burned to CD-ROM for archival and for shipping to the external marker. We found the system extremely reliable, and resilient to operator error or equipment failure. The auto-saving aspects of OpenOffice allowed us to retrieve, virtually intact, the work of two students who had this kind of problem.

The results of the pilot were highly encouraging, with all candidates submitting completed scripts in digital format. These were marked externally, and the results returned as a spreadsheet. Anecdotal evidence from student remarks indicated they did not like the timed nature of the assessment (many carry-home assignments do not have this limitation), but did appreciate the opportunity to use a computer for a formal test "because it makes sense". Compared to a related assignment (on teaching through animation) in the unit, the computer-based examination was slightly harder, with fewer students getting higher grades (see Table 1).

**Table 1: Numbers of students with each grade**

	Reverse cumulative grades on computer-based examination	Reverse cumulative grades on related assignment
Fail	100%	100%
Pass	100%	96%
Credit	86%	84%
Distinction	34%	58%
High distinction	12%	23%

These figures are not in themselves evidence for the efficacy of the computer-based examination system, but assure us that it is sufficiently resilient to give results comparable with other assessment techniques. Since this was a new unit, comparison with previous years was not possible; nor was a split cohort using different testing regimes possible on equity grounds. A future use of the technique might be ethically approved if students self-selected between different modes of taking the same assessment.

## Discussion

### *Text creation in examinations*

This study illustrates the strategic problem posed by formal assessment methodologies; and hence the importance of providing a reliable and resilient system whereby undergraduate assessment can use the computer as a principal text-creation tool in examinations. To do so will allow other areas of

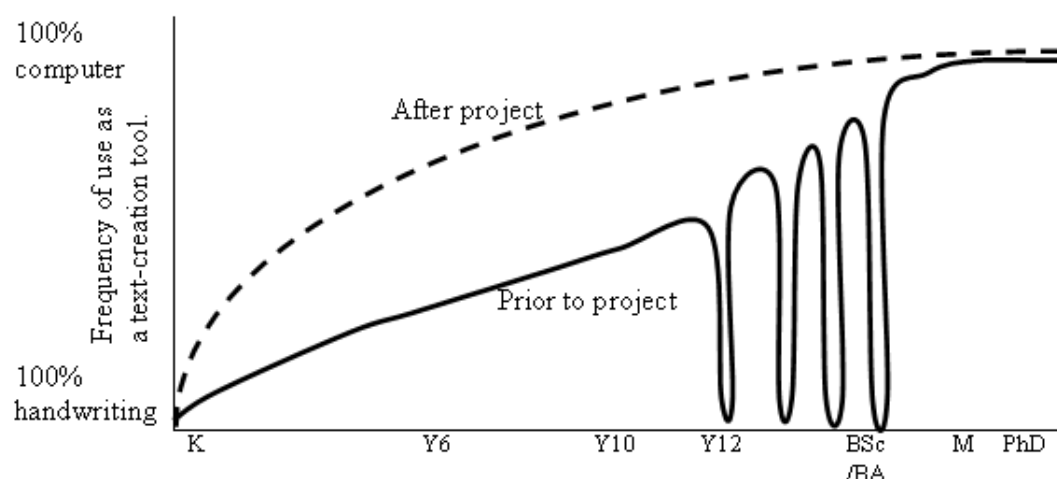
education to use similar methods. Conversely, so long as first degrees are awarded primarily on the basis of hand-written exams, this will remain the de-facto standard for all areas of schooling.

Whilst some US law schools administer examinations taken on tablet computers, they still distribute the questions on paper (Augustine-Adams et al 2001). This project eliminated paper completely from the assessment activity, yet provided potential opportunities for candidates to use their own laptops, university desktops or a suite of special-purpose computers.

This pilot study established the proof-of-concept for an open-source system to replace printed examination papers by live operating system CD-ROMs. In a non-networked environment, collusion was suppressed, but this did make the collation of completed scripts more difficult. In the future, the excision of networking drivers from the CD-ROM may be adjusted and allow internal submission to a local server. This could be controlled by physically unplugging from the local router any cable leading elsewhere.

### *Cultural Transformation*

Figure 2 illustrates the anticipated effect of tertiary eAssessment on student writing behaviour over the course of an academic learning pathway. The solid line below represents anecdotal evidence and the literature. The anecdotal evidence suggests that pupils in schools are increasingly allowed to submit assignments and homework in printed format as they progress through the school system. Talks with teachers in the primary sector emphasise the need for pupils to concentrate on and develop good handwriting skills at an early age, hence the gradual increase in machine-mediated text production. Externally moderated examinations in Years 11 and 12 and undergraduate degrees are currently required to be completed using handwriting – hence the annual dips. Currently, choice of a computer as a writing tool is impeded by high stakes examinations from Year 12 (age 18) to completion of a Bachelor's degree. If we are successful in removing this barrier to innovation, students will be able to choose the appropriate writing tool for any given task more freely. This will facilitate curriculum transformation through ICT across the breadth of diverse subject areas taught in the period of compulsory schooling.



**Figure 2: Anticipated student writing behaviour**

## Conclusion

Future activities could involve the comparison of schools which feed students into the same university. Laptop schools or those with a significant investment in ICT may be more likely to transform their curriculum delivery if the tertiary institution exhibits a receptive tendency to students versed in that medium.

Some supervised forms of written assessment can be undertaken in an ICT-based environment. This could be used for performance assessments, knowledge assessment, some professional skills assessments and to facilitate essay, short or long answer written tests. Where these tests are normally conducted in an examinations hall, the venue could be moved for the purposes of the trial to standard computer laboratory(ies) when student numbers are sufficiently small. Therefore some University examinations will be taken by students using computers instead of by handwriting.

The benefits for students include:

- An context for assessment similar to the context for learning for most students (who are IT-savvy and access many units through on-line materials)
- The capacity to perform changes and re-organise written replies at any time up until the end of the examination without crossing out or attempted erasure
- Fewer students with disabilities would need separate conditions, leading to inclusivity of practice.

For University staff, the benefits include:

- For examination supervisors, a simple way to verify the correct environment is in use (current requirements with respect to checking the technical capabilities of electronic calculators can cause concern);
- Easier marking for examiners: the digital scripts can be marked on-screen or from printouts – no more guessing what the student was attempting to hand-write;
- Opportunities to streamline administration when the scripts and marks are retained in a single digital environment, by eliminating transcription errors.

The characteristics of an ICT-based examination system will include:

- Portability – it should be possible to set it up using almost any available equipment
- Equity – it should be accessible to a wide range to students
- Familiarity – students should have every opportunity to practice essential skills in this environment



- Technical capacity – it should not limit students creativity or expression
- Archivability – the environment should produce material which will be accessible in future years
- Inviolable – students should not be able to alter the environment to gain an unfair advantage. (Fluck, 2004b)

By replacing formal tertiary examination papers by CD-ROMs, printing costs have been saved and marking expedited. It is expected that this cost saving and gain in efficiency will be matched by greater satisfaction from students who are rarely required to write by hand, except in high-stakes testing.

As tertiary institutions replace hand-written examinations by supervised computer-based activities, Australian schools will be empowered to use information and communication technology (ICT) in more challenging and transformational ways which reflect the realities of modern life.

## References

- AICTEC (2006) *Learner identity management framework project: Framework report (v3.0)*. Convergence e-Business Solutions Pty Ltd. Online at <http://www.aitect.edu.au/aitect/go/engineName/filemanager/pid/220/Final%20Report%20March%202006.pdf> on 11th January 2007.
- Ashton, R. and Thomas, H. (2006) Bridging the gap between assessment, learning and teaching. *Proceedings of 10<sup>th</sup> International Computer Assisted Assessment Conference, 4-5 July* Loughborough University, Leicestershire.
- 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
- Blackboard (2007) *Exambient*. Online at <http://www.blackboard.com> on 11<sup>th</sup> Mat 2007.
- Chapman, G. (2006) Acceptance and usage of e-assessment for UK awarding bodies – a research study. *Proceedings of 10<sup>th</sup> International Computer Assisted Assessment Conference, 4-5 July* Loughborough University, Leicestershire.
- Downes, T, Fluck, A, Gibbons, P, Leonard, R, Matthews, C, Oliver, R, Vickers, M, Williams, M, (2002) *Making Better Connections*, Commonwealth Department of Education, Science and Training.
- Enerson, Diane M. (1997) *Report on using computers as an aid to teaching and learning: Learning from existing practice*. Center for Excellence in Learning and Teaching, The Pennsylvania State University. Retrieved September 22, 2004 from <http://www.psu.edu/celt/computer.html>
- Fluck, A (2003) *Integration or Transformation? A cross-national study of ICT in school education*. University of Tasmania: PhD Thesis. Online at <http://eprints.utas.edu.au/232/02/02whole.pdf> on 9 February 2007.
- Fluck, A, (2004a) 'Government sponsored open source software for school education', *Proceedings of the IFIP 18th World Computer Congress - TC3/TC9 1st Conference on the History of Computing in Education*, Toulouse, France, 27-45
- Fluck, A, (2004b) 'Can students use computers for formal examinations?', *Proceedings – Teaching Matters conference*. University of Tasmania: Hobart, Tasmania
- Fluck, A, (2005) 'The realities of transforming education through ICT', invited keynote at *Global Project Based Learning Forum and Exhibition, Kaohsiung, Taiwan*, 29-35 (Eng)
- Hawkridge, David (1989) Machine-mediated learning in third-world schools. *Machine-Mediated Learning*, 3, 319-328
- Hibbitts, Bernard J. (1996, Sept) The Interface is the Message. *Wired*.
- Knopper, Klaus (2004) *Knoppix: What is Knoppix?* Retrieved December 15, 2004 from <http://www.knopper.net/knoppix/index-en.html>.
- Maslen, Geoff (16-22 June, 2004) Type-casting of students for VCE: exams online. *Campus Review*
- McGill, Lou (2006) 'Overview of JISC Assessment Activities'; in *Proceedings of 10<sup>th</sup> International Computer Assisted Assessment Conference, 4-5 July* Loughborough University, Leicestershire.

- Mogey, Nora (2006) Essay Exams and Tablet Computers – Trying to Make the Pill More Palatable; in *Proceedings of 10<sup>th</sup> International Computer Assisted Assessment Conference, 4-5 July* Loughborough University, Leicestershire.
- Office of Government Commerce (2004) *Government open source software trials final report*. Available online [www.ogc.gov.uk/index.asp?id=2190](http://www.ogc.gov.uk/index.asp?id=2190) accessed 31 October 2006.
- Pescaru, Dan and Holotescu, Carmen (2002) Authentication in an online learning environment: A case study. *Proceedings RoEduNet Conference 2002*. Online at <http://www.timsoft.ro/ejournal/roedunet2.html> on 4th July 2003.
- Reynolds, D; Treharne, D and Tripp, H. (2003) ICT – the hopes and the reality. *British Journal of Educational Technology* 34(2) 151-167.
- SecureExam (2006) *Our product suite*. <http://www.softwaresecure.com/suite.htm> on 11th May 2007.
- Tinker, R. (2000) Ice machines, steamboats, and education: structural change and educational technologies. Paper presented at *The Secretary's Conference on Educational Technology*, September 11-12. Available online <http://www.ed.gov/rschstat/eval/tech/techconf00/tinkerpaper.pdf> accessed 31 October 2006
- UK Passport Office (2004) *The UKPS biometrics enrolment trial*. Retrieved December 15, 2005 from <http://www.passport.gov.uk/docGallery/17.PDF>.
- Winkley, John and Osborne, Che (2006) Developments in On-Screen Assessment: Design for Examinations; in *Proceedings of International Computer Assisted Assessment Conference, 4-5 July* Loughborough University, Leicestershire.
- Zymaris, Con (2004) *OpenOffice.org*. Bondi Junction, NSW, Australia: Derwent Howard.



**ASSESSMENT AND LEARNING:  
IS ASSESSMENT AN  
AFTERTHOUGHT OR IS IT AT THE  
HEART OF THE LEARNING  
PROCESS?**

**Dr Bob Gomersall**



# Assessment and Learning: Is Assessment an Afterthought or is it at the Heart of the Learning Process?

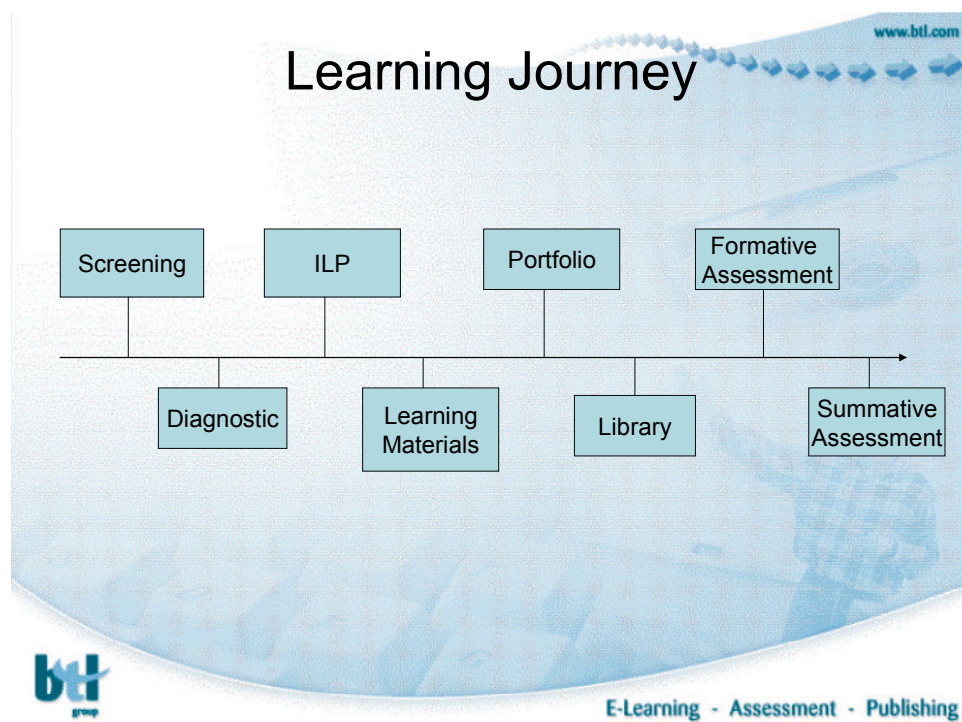
Dr Bob Gomersall, Chairman, BTL Group Ltd. [www.btl.com](http://www.btl.com)  
bob.gomersall@btl.com

## Abstract

An approach to learning is described which is built on techniques developed for on screen assessment and formative assessment. It aims to provide a high level of motivation, immediate student centred feedback and a high level of learner control. The technology (known as btl engage™) can be applied to any area in which on screen assessment material is already available, extending it into areas such as revision, interactive worksheets and e-learning.

## Background

The traditional Learning Journey consists of a series of learning experiences followed at the end of the process by a summative assessment. Some typical examples are set out below.



The format and style of the final assessment will drive the learning styles throughout the process. If the summative assessment is paper based, as is usually the case, the learning will reflect this. If the summative assessment is screen based then it is reasonable to expect that the learning styles will change, becoming more screen based themselves, but in addition there is no reason to assume that the traditional linear Learning Journey will remain intact.

In addition the Learning Journey has traditionally been driven by the teacher. Furthermore, a growing interest in formative assessment (Black and Wiliam (1998)) has led to this being seen as one of the key learning experiences in the Learning Journey, with much of the feedback taking place through a teacher or via scores and statistics (Mann and Glasfurd-Brown (2006)) . Whilst progress in this area is seen as very significant, it is hardly the self-regulating route of Yorke (2003) or the student centred route that is the natural consequence of e-learning and e-assessment.

Two innovations are therefore likely to lead to a re-shaping of the Learning Journey – on screen assessment leading in turn to more on screen student centred learning. The aim of this paper is to show how a student centred approach to formative assessment can re-shape the Learning Journey. The pragmatic reasoning behind the approach is set out, along with some practical actions and early results.

### **Current Position in e-assessment and e-learning**

There has been a rapid development in the use of on-screen testing, with large numbers of candidates taking tests in this form. In some areas, such as Skills for Life testing in the UK, the majority of tests are already on screen. A number of lessons have already emerged (see for example Osborne C and Winkley J (2006)).

- The majority of candidates prefer on screen tests.
- The results are better than those of candidates using paper based tests, although the reasons are not well understood.
- Where work is automarked candidates appreciate the rapid feedback of results.
- The administrative benefits offer greater opportunities for formative, screening and diagnostic assessments.
- The separation between assessment and learning is likely to be less distinct in the e-world than it is when using paper based systems.

On the e-learning side the current position has been very carefully set out by Clarke (2004) and Clark and Mayer (2002). Clarke's book provides a comprehensive survey of all aspects of e-learning. Clark and Mayer conduct a very careful analysis of what does and does not work in an e-learning context building their arguments on a solid research base. Their arguments are



facilitated by some simple classifications of e-learning approaches, which can in principle be applied to any learning context (not just in e-learning).

Of particular interest in the current context is Clark and Mayer's classification of the three types of e-learning, as shown in the table.

Type	Description
Receptive	Show and Tell
Directive	Tell and Do
Guided Discovery	Problem Solving

The short descriptions – “show and tell” etc – are a shorthand for describing the interaction of the learner with the environment, and this will be developed further below.

Works of the above type are extremely useful to e-learning developers. They do exactly what they set out to do - describe how things are done *now* rather than how they *might be* done in the future. To compare this with the early days of the railway, when trucks were pulled by horses, these e-learning books provide excellent manuals for the maintenance of the railway line, the grooming of the horse and the oiling of the truck's wheels.

In fact Clark and Mayer also briefly look into the future, and try to discern the shape of the steam engine. This paper attempts to build on some of those ideas.

## Theory

The obstacle we face is that we have no adequate framework for our thinking to allow us to predict the outcome of any given course of action. Since this is the most basic requirement of a “theory”, we have to conclude that we do not have an adequate theoretical base – notwithstanding the work of Clark and Mayer which provides an excellent empirical base founded on psychological research. Indeed it may well be that given the complexity of the situation no theory in the scientific sense of the word will be possible for a long time yet. On the other hand, if we are to make progress there is a need for some pragmatic guidelines, and the purpose of this paper is to suggest how these might be put together.

The aim is to create a framework which can guide our thinking, allow us to see traditional approaches in perspective and indicate a way of moving forward so that the predictions of what will and will not work can be tested against the actual outcomes.

The discussion will be structured in three parts as follows.

- Interactions of the learner with the environment
- Thought processes of the learner
- Routes through the learning materials

The approach will not draw heavily on psychological theory but rather on simple pragmatic concepts which have proved useful to the author in generating real solutions that people are willing to pay for.

## Learning Elements

Most of the interactions the learner has with the environment can be summarized in a single word – **Tell, Show, Guide, Try/Do, and Assess**. Clearly this list is not comprehensive, since it should also include touch, taste and smell, but in the context of paper based and on screen learning and assessment these are less relevant – for now!). The aim of this section is to argue that each of these actions can be described as a Learning Element or building block, from which a learning experience – and in particular an e-learning experience – can be built up.

Taking the three types of e-learning described by Clark and Mayer, we can see how the Learning Elements are assembled in those particular instances.

Receptive = Tell + Show

Directive = Show + Do

Guided Investigation = Guide + Do

Each of the methods consists of a pair of Learning Elements. A little thought shows that many other combinations are possible, and indeed correspond to well understood teaching and learning styles. Of course a learning experience may consist of one Learning Element or many. Effective teachers have always invoked the full range of Learning Elements, subject to the limitations of the classroom. In Table 2 there is an assessment of the level of usage of the different Learning Elements in traditional teaching, along with a summary of some of the new opportunities offered by e-learning.

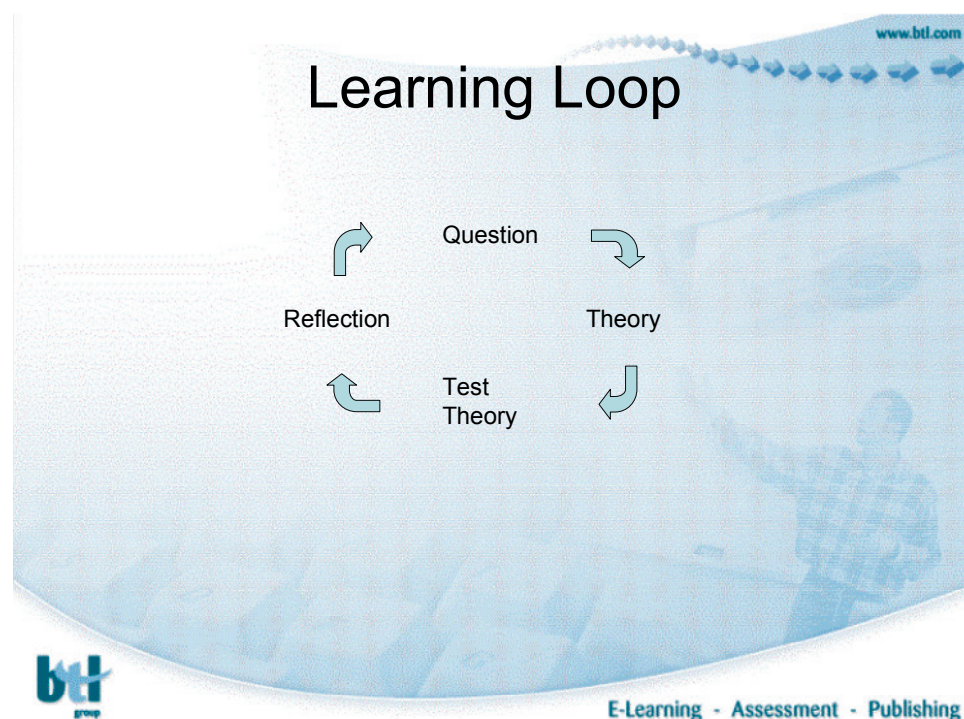
**Table 2: Use of Learning Elements - Traditional and e-learning**

	<b>Traditional Learning</b>	<b>e-learning Opportunity</b>
<b>Tell</b>	Very large	Opportunity to provide a more consistent quality
<b>Show</b>	Large	Opportunity to improve quality using colour, animations, video, images etc
<b>Guide</b>	Modest (limited by teacher time)	Major opportunity to provide feedback
<b>Try/Do</b>	Modest	Vast opportunity - Feedback can add motivation
<b>Assess</b>	Large	Major opportunities - instant feedback, simpler administration, opportunities with formative assessment, screening and diagnostic.

## Thought Processes of the Learner

The previous section deals with the interaction between learner and environment. This section considers a pragmatic way of looking at the thought processes of the learner. This is of course the subject of a vast amount of literature. However, in order to remain faithful to the initial aim of pragmatism, the approach in this section is to set out some simple ideas for framing our thoughts.

The basis of this is the proposition that learning is a (hopefully) streamlined version of what happens when someone learns something for the first time i.e. when it is discovered. This is well understood and has been described by many authors such as Popper and Kuhn. Just as *discovery* can be described as acquiring knowledge or understanding that was previously unknown, so *learning* can be described as acquiring knowledge or understanding that is unknown to the learner, but is already known or understood by others. An equivalent view is that for an individual any learning represents discovery for the first time. Handy (1989) gives a simple summary of this approach, drawing on the ideas of Kolb, which is encapsulated in the “learning loop”.



Referring to the diagram above, the following describes the key features.

1. **Question:** The learning is initiated by a problem, a question, a puzzle, a challenge to be met or a dilemma to be resolved.
2. **Theory:** The learner then formulates a theory of how to address the problem and arrive at an answer. This may be very simple and held in the head or it may be very complex and require the use of

additional external tools such as written language or mathematics. The term “theory” is used in its widest sense, from a loose hypothesis to a well established scientific theory.

3. Test Theory: The predictions of the theory are then tested against experience and existing knowledge. These may accord with existing experience (the “expected” result) or they may not.
4. Reflection: If the results of the theory are as expected then the learner may move on to a new question or problem. If the results are unexpected then the learner will need to re-visit the question and re-formulate the theory.

The above is a simple summary of the so-called “scientific process”, but in practice it is the method by which all reliable knowledge is gained. In science the predictions of theory (initially known as a hypothesis) are tested against experiment, and if the predictions do not accord with the experimental outcome then the theory has to be re-visited and amended. In principle it only requires one type of experiment to disagree with the predictions of a theory (“falsification”) for the theory to be abandoned. In practice of course it will require a lot of checking and re-checking of experiments before any such thing happens, especially in the case of theories which are at the heart of our scientific culture, but that is still the way it works. The works of Kuhn and Popper deal with this area in great detail.

The point about this is that all learners – if they really learning - are going round and round this loop, being driven each time by a problem or question. This is an internal process and it goes on all the time. The problem a teacher faces is how to direct this learning in the way desired rather than the way the student wishes (which may be more concerned with something entirely different – and probably more interesting - such as getting a girlfriend or improving performance in a computer game).

This gives us an insight into the problem of teaching – namely to persuade the student to move round the learning loop. Many good teachers usually start by outlining the problem before embarking on an explanation. In so doing they are seeking to drive the student round the loop. Whether it works is a different matter. The student may write notes, but actually think about how to get to the dining room before the queue gets too long – a much more pressing problem. Some teachers never answer questions except with another question. In the right hands this is another very effective technique, which repeatedly drives the student round the learning loop. Whole courses have been devised around this concept of a Socratic Dialogue (see for example [www.physics.indiana/~sdi](http://www.physics.indiana/~sdi)), and most people have encountered teachers who have adopted this approach to a greater or lesser extent.

This then leads to a hypothesis about the learning process: *Learning is most effective when it follows as closely as possible the discovery route or learning loop.*

It follows that *learning is most effective when it is led by a problem or a question*. This is the opposite of the approach taken traditionally, in which learning material precedes assessment. Broadly speaking the teacher explains the material, the students learn it and are then tested on it. There is sometimes a cursory mention of the question being addressed by the particular knowledge being imparted, but this is seldom the centerpiece of the activity.

Leading with a question was the basis of the Socratic Dialogue approach and more recently underpinned the discovery learning approach adopted in much of Nuffield Science. In practice the approach to the latter had to be significantly modified because it was not easy to constrain the problem sufficiently in a practical context to avoid huge wastes of time – but handled well the approach did have a real impact on teaching and learning which permeates science teaching of all types today.

In the context of e-learning, two points need making. Firstly it is clear that e-learning can offer a new approach to the learning loop, driven by the learner rather than the teacher. Secondly, and rather more specifically, the use of simulation offers the opportunity to constrain a problem much more precisely than was ever possible with practical work. (Adding this to the other opportunities presented by simulations hints at the wider possibilities offered by this approach (Thomas et al (2005))). In addition the range of applicability is much wider, covering all subjects and many areas which are otherwise impossible as a result of being too large, too small, too expensive, too dangerous, too complex or – significantly – too abstract.

Finally, feedback in the e-learning context can in principle take place at precisely the point at which it is required – at the point of cognitive conflict. It should therefore be possible to highlight the *location* of an error without giving an explanation. This would be a significant step forward because it would face the student with a question or problem at precisely the right point, and avoid the need to plough through large amounts of correct work in order to “find a mistake”. A method for achieving this is described below, following a summary of the points outlined so far.

## Summary so far

So far the following points of view have been advanced.

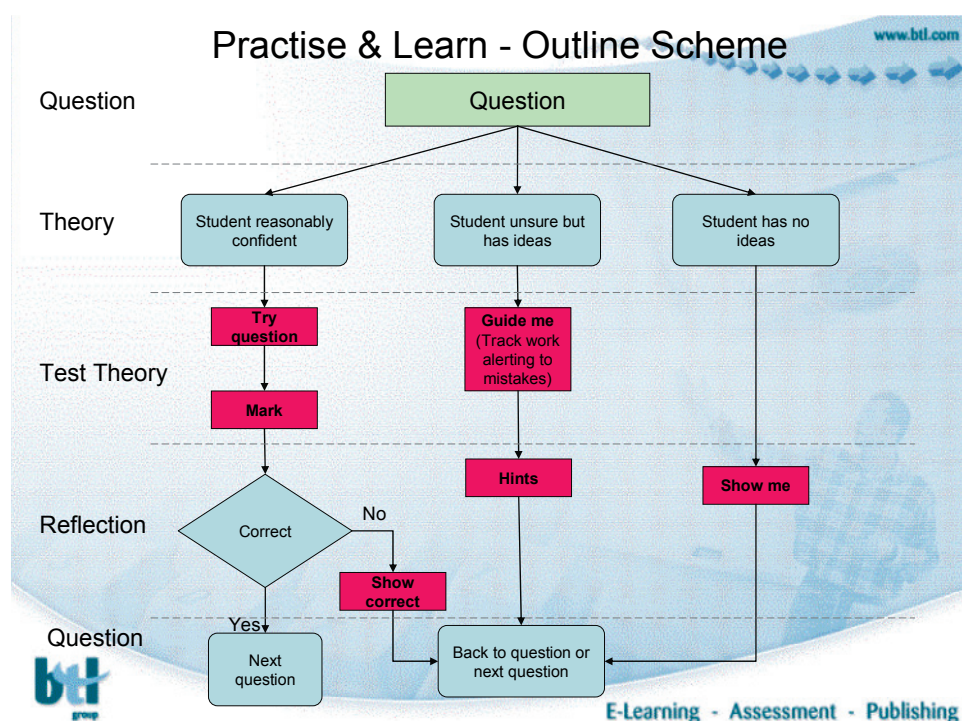
1. The learning process involves a number of types of interaction with the environment which can be characterized as *Learning Elements*. These include **Tell, Show, Guide, Try/Do and Assess**.
2. Learning is most effective when it follows as closely as possible the discovery route or *learning loop*; when it is initiated by the learner and is an *internal* (rather than an imposed) process; when it is *led by a problem or question*; and when there is feedback at the level of the individual question or part question.

3. The most important feature of feedback is to highlight the *location* of an error rather than to supply a correction or explanation.

On the basis of these hypotheses, the e-learning challenge is therefore to devise a means of combining the Learning Elements, and a route through them, which motivates the learner to follow the learning loop as effectively as possible. This is addressed in the next section. Initial trials look very promising, due in part to the opportunities offered by immediate feedback, but also as a result of the inherently student-centred nature of the approach.

### \* btl engage™

The technology known as btl engage has been designed by the author with the aim of taking advantage of the conclusions set out earlier. The outline scheme is set out in the following diagram.



Learning starts with a question or problem, as shown in the diagram. This is important not only from an educational point of view, but also from a practical perspective, because it constrains the number of routes that need to be made available to the learner to manageable proportions. The question itself may in fact have been chosen by the learner or a teacher, depending on the circumstances, and indeed the approach to selecting questions opens up a whole range of new opportunities for learning (see later).

The learner is initially in one of three broad states of mind – reasonably confident, unsure but has ideas or has no ideas at all. The three available routes are designed to meet these three different situations. A reasonably confident learner can take the **Try/Do** option and attempt to answer the question. When complete this can be automatically marked, with an option to

be given the correct answer if an error is made. At the other extreme, the learner who has no ideas can simply opt for **Show** and can be taken stepwise through a model solution.

The unsure student has an intermediate option – **Guide**. In many ways this is the most interesting because it attempts to imitate the situation of a teacher looking over the shoulder of the learner, pointing out mistakes and possibly dropping hints. The computer tracks the work of the learner and at each step an indication is given that it is correct or incorrect. As a result, the learner can proceed confident in the knowledge that they are on the right track. As soon as a mistake is made it is flagged up and the learner can focus all their mental energy on solving the problem in hand, rather than, as is all too often the case traditionally, devoting a lot of effort into locating the error in the first place. Finally hints can be made available and even additional tutorial material.

The combined effect of the above scheme is to provide the equivalent of a series of questions, with worked answers available for every one, with the opportunity to try any without being told the answer and yet immediately check at the end, and finally the opportunity to have work checked on a real time basis without feeling any of the pressure normally associated with a teacher looking over the shoulder.

In summary, btl engage aims to provide a framework for student centred learning which draws on the ideas set out above. In particular it addresses the following issues.

- It sets out to provide thorough coverage of three Learning Elements – **Show**, **Guide** and **Try/Do** – along with a simple development route to involving all Learning Elements.
- It aims to follow as closely as possible the idea of the Learning Loop, and most importantly it is led by a problem or question.
- Feedback is integral to the process, and a key feature of **Guide** is the ability to highlight the location of an error in real time.

As described above, the route through the materials can be determined by the learner. In practice there is nothing to stop a teacher using the materials in a more restricted way, providing a route through the learning content.

### **Applications of btl engage™**

The technology can be applied in a number of different ways.

Interactive questions: These may be set out rather like textbook questions, classified by type and graded according to the level of difficulty. Students could be directed to the best starting point by a teacher, but in practice it may well be more effective for students to determine their own starting point, and their own pace through the material.

Interactive worksheets: These would be similar to interactive questions.



Interactive revision materials: These may involve questions on a wide variety of subject matter, with little connection between one question and another. As a revision tool it would be potentially very powerful.

Learning materials: Carefully selected questions could in principle guide the learner through any learning materials. In practice, it may well be that there is no difference between this and interactive questions – simply a much more comprehensive set of questions. Indeed it is possible that the traditional approach to teaching in which the content is explained and then the learners tested may indeed have no place at all in the student centred e-learning world.

In practice it is likely that the traditional distinctions between questions, worksheets, revision materials and learning materials will become increasingly blurred. This in turn suggests that the linear Learning Journey could well be replaced by a question led screen based “socratic dialogue” in which the student has much more direct control over the learning.

A sample screenshots in which the technology has been embedded in a learning package (Practise and Learn) is shown below. Each entry by the learner is marked as soon as it has been entered, providing instant feedback.

**Btl engage™ – Guide Me** [www.btl.com](http://www.btl.com)

The screenshot displays the 'Practise & Learn' window within the Btl engage application. The title bar includes 'File View Control Help'. The window title is 'Practise & Learn' with 'Help' and 'Quit' buttons. The main content area is titled 'Add and subtract fractions - section b' and includes 'Return to skill area' and 'Return to menu' links. The problem instruction is: 'Change  $\frac{2}{5}$  and  $\frac{1}{3}$  into equivalent fractions with the same denominator.' Below this, there are visual aids: a bar model for  $\frac{1}{3}$  (a bar divided into 3 equal parts, with 1 part shaded blue) and a bar model for  $\frac{2}{5}$  (a bar divided into 5 equal parts, with 2 parts shaded blue). Further down, there are two more bar models for the same fractions, each followed by an equals sign and a grid of 15 small squares (3 rows by 5 columns). The first grid for  $\frac{1}{3}$  has 3 squares shaded blue, and the second grid for  $\frac{2}{5}$  has 6 squares shaded blue. Below these grids, the fractions  $\frac{1}{3} = \frac{3}{15}$  and  $\frac{2}{5} = \frac{2}{15}$  are shown, with the numerators 3 and 2 entered in red boxes. A 'Mark' button is located to the right of the second fraction. At the bottom of the window, there are buttons for 'Try question', 'Guide me', and 'Show me'. To the right of these buttons are 'Next question' and 'Previous question' buttons, and a numeric keypad with digits 1 through 10. The Btl group logo is in the bottom left corner, and the text 'E-Learning - Assessment - Publishing' is in the bottom right corner.



## **Predictions, Benefits and Issues**

In principle many of the proposed theoretical criteria for improvements in learning are met by btl engage<sup>TM</sup>. If the theory is to have any value then these predictions need further testing.

To date a simple qualitative trial has been conducted with a small group of Year 10 students (Brumfitt M (2006)). The content was aimed at revision of fractions, an area that is notoriously difficult. The students provided anonymous feedback through a simple questionnaire, and in summary the following benefits were identified by the learners.

- The students liked the look of the tool and found navigation “nice and easy”.
- They found that the “Guide Me” and “Show Me” tools were interesting to work with, and that they were particularly useful with questions that they were not very confident in answering. One individual comment was that “‘Guide Me’ was extremely useful because it gave you a chance to still answer the question off your own back with a slight nudge in the right direction, whereas in a text book, although the workings out are shown, they give the answer as well – preventing you from answering the question”.
- The potential for revision was highlighted particularly.
- It was pointed out that when using textbooks to revise it was necessary to jump around constantly from chapter to chapter, which could be time consuming, whereas using the tool everything they needed was literally a few clicks away. This was found to be a much more practical way to revise.
- The students genuinely enjoyed it and were impressed with the functionalities.

Feedback from the teacher of the group and a student teacher was enthusiastic but otherwise broadly similar.

The issues that emerged were as follows.

- The extent to which learners should have freedom of navigation through the menu of questions needs further work.
- A timing function should be considered.
- Availability to learners at home as well as at school is a key consideration.

Clearly further thorough trials are required to establish just how effective the proposed methods are, and indeed to establish whether the learning process is significantly speeded up, whether levels of motivation and engagement improved, or the process shows some other practical measurable benefits. It

is predicted that there will be significant benefits, and the early work is promising - but the basic proposition remains to be properly tested.

\* Subject of a patent application in the UK and US.

## References

- Black, P. and Wiliam, D. (1998). *Inside the Black Box: Raising Standards through classroom assessment*. Department of Education and Professional Studies, Kings College, London.
- Brumfitt, M. (2006). Internal BTL Report
- Clarke, A. (2004). *e-Learning Skills*, Macmillan
- Clark, R.C. and Meyer, R.E. (2002). *e-Learning and the Science of Instruction*, Pfeiffer
- Handy, C. (1989). *The Age of Unreason*, Business Books Ltd.
- Mann, H., and Glasfurd-Brown, G. (2006). Learning from Assessment: Evaluating the benefits of DALI (Diagnostic Assessment Learning Interface), *Proceedings of the 10<sup>th</sup> International CAA Conference*, 2006.
- Osborne, C. and Winkley, J. (2006). Developments in On-Screen Assessment Design for Examinations, *Proceedings of the Tenth International CAA Conference*, 2006
- Thomas, R., Ashton, H., Beevers, C., Edwards, D. and Milligan, C. (2005). Cost effective use of simulations in online assessment. *9<sup>th</sup> International Computer Aided Assessment Conference*, Loughborough University, Loughborough.
- Yorke, M. (2003). Formative assessment in higher education: Moves toward theory and the enhancement of pedagogic practice. *Higher Education*, 45(4), 477-501.



**A NEW METHOD FOR PARSING  
STUDENT TEXT TO SUPPORT  
COMPUTER-ASSISTED  
ASSESSMENT OF FREE TEXT  
ANSWERS**

**Elizabeth Guest and Sally Brown**



# **A New Method for Parsing Student Text to Support Computer-Assisted Assessment of Free Text Answers**

Elizabeth Guest and Sally Brown, Leeds Metropolitan University

## **Abstract**

Due to current trends in staff-student ratios, the assessment burden on staff will increase unless either students are assessed less, or alternative approaches are used. Much research and effort has been aimed at automated assessment but to date the most reliable method is to use variations of multiple choice questions. However, it is hard and time consuming to design sets of questions that foster deep learning. Although methods for assessing free text answers have been proposed, these are not very reliable because they either involve pattern matching or the analysis of frequencies in a “bag of words”.

The first step towards automatic marking of free text answers by comparing the meaning of student answers with a single model answer is to parse the student work. However, because not all students are good at writing grammatically correct English, it is vital that any parsing algorithm can handle ungrammatical text. In this paper, we present preliminary results of using a relatively new linguistic theory, Role and Reference Grammar, to parse student texts and show that ungrammatical sentences can be parsed.

## **Introduction**

In the current climate of increasing student numbers and decreased funding per student in many HEIs internationally, it is necessary to find economies of scale in teaching and supporting undergraduate students. Economies of scale are possible to a certain extent for lectures and tutorials, but this is less possible for assessment. As staff student ratios decrease, the assessment burden on staff will increase unless alternative approaches are used.

One solution to this dilemma is to seek ways to mark student work automatically. This is being done at present using variations on multiple choice questions, with a variety of innovative question types that test learning beyond simple recall. If designed correctly, these kinds of tests can provide students with immediate feedback on how well they are doing and can provide valuable formative pointers for further learning. Extensive evidence demonstrates that increased formative assessment can impact positively on student learning and retention (Sadler, 1989) (Sadler, 1998) (Rust, 2002) (Sambell and Hubbard, 2004) (Yorke, 2001). However, it can be difficult to design if we want to make it truly an integral part of learning and if we want to

avoid encouraging inappropriate student behaviour, such as random guessing of answers.

Considerable work has been undertaken in recent years to investigate and implement approaches to CAA that foster deep learning (Beevers et al., 1989) (Brown et al., 1999), but significant advances still remain to be made. Some have argued that it is currently possible to assess essays by automatic means but we remain unconvinced. However, it would be very helpful if it were possible to automatically mark short free text answers using CAA approaches, thus reducing the drudgery for markers. This would allow more scope in the setting of questions and would give students more opportunity to show what they understand and can do. Much research has been aimed at this question, but this generally either involves pattern matching (Sukkarieh et al., 2003) (Sukkarieh et al., 2004) or latent semantic analysis (Wiemer-Hastings, 2001) (Landauer et al., 1997), or a combination of these (Pérez and Alfonsa, 2005). These methods work to a certain extent, but because they are not based on the meaning of the text, they are quite easy to fool. For instance latent semantic analysis can be fooled by writing down the right kinds of words in any order. The problem with current approaches to pattern matching on the other hand, is that if the student writes down a correct answer in a different way, it will be marked wrong.

Our innovative approach is based on the grammatical tradition of parsing, that is breaking down language into its functional components like verbs, nouns and adverbs. Role and Reference Grammar (RRG) (Van Valin and LaPolla, 1997) (Van Valin, 2005) is a relatively new linguistic theory that majors on predicates and their arguments. It separates the most vital parts of the sentence from the modifiers (adverbs, adjectives, auxiliaries, and articles). This means that the core meaning can be extracted first and then the modifiers fitted in at a later stage. As long as the arguments and the verbs are in the correct order for English (subject verb object) then the sentence can be understood. It doesn't matter if (for example) Chinese students forget the articles, the sentence can still be parsed and the meaning extracted. The core meaning of the sentence is extracted via the use of templates. This makes it easier to extract the important parts of the meaning of the sentence: we just need to identify the predicate and the arguments which are clearly labelled branches within the templates.

In this work we describe a method for using the RRG paradigm for parsing student texts, which do not have to be grammatically correct. This work can be used as a pre-processing step to those methods that use latent semantic analysis or pattern matching. There is evidence to suggest that latent semantic analysis gives better results when the subject, verb, and object of the sentence is used rather than an unstructured "bag of words" (Wiemer-Hastings, 2001). Our method will provide a mechanism for extracting some structure. If structure can be extracted, then this structure can also be passed to a pattern matcher, which will decrease the number of possibilities that have to be included. This method will also enable accurate marking of ungrammatical sentences.

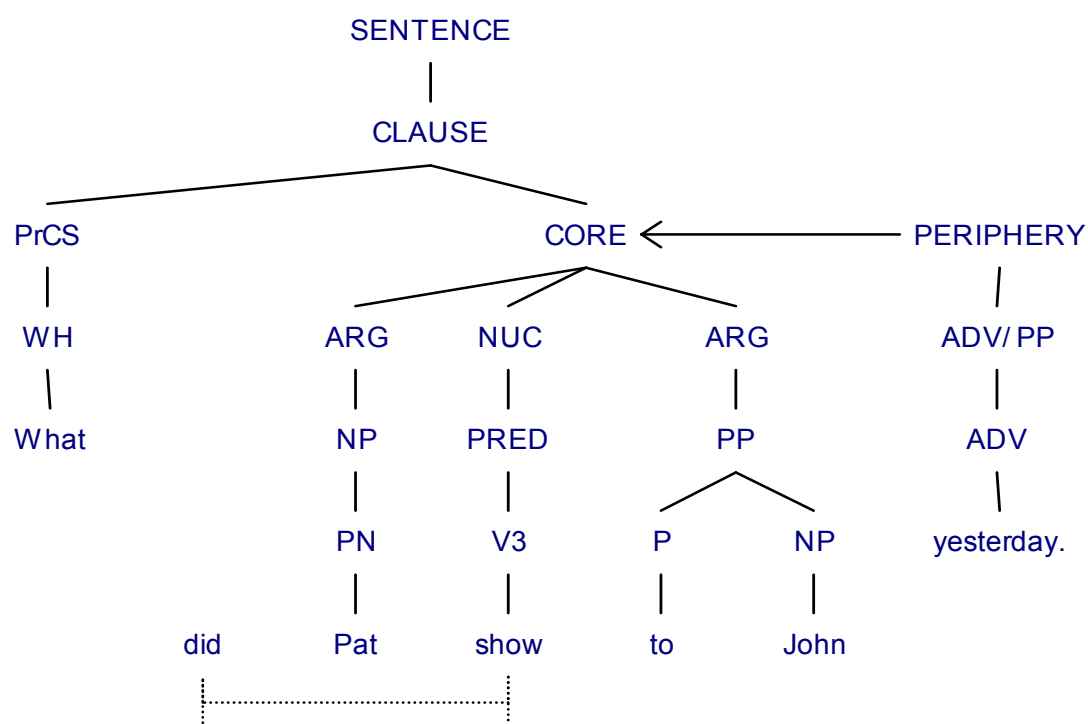


## Parsing for Role and Reference Grammar

Role and Reference Grammar (RRG) (Van Valin and LaPolla, 1997) (Van Valin, 2005) was developed as a result of asking the question “What would a linguistic theory look like if it was based on Lakota and Tagalog rather than English?”. The result is a theory that is suited to describe a huge range of languages, including English. Of all the linguistic theories, it is most closely related to functional grammar, but there are important differences.

Role and Reference Grammar posits algorithms to go from syntax to semantics and semantics to syntax. The main contribution is the use of parsing templates and the notion of the CORE. A CORE consists of a predicate (generally a verb) and (normally) a number of arguments. It must have a predicate. Everything else is built around one or more COREs. Simple sentences contain a single CORE; complex sentences contain several COREs.

The fact that RRG focuses on COREs, means that the semantics is relatively easy to extract from a parse tree. You just have to look for the PRED, and ARG branches of the CORE to obtain the predicate (PRED) and the arguments (ARG). Who did what to whom will depend either on the ordering of the ARG branches (in the case of English), or on their cases, or both.



**Figure 1: Example RRG parse tree.**

An example of an RRG parse tree is given in figure 1. Notice that in this example, the word “did” does not feature in the parse tree, but it is linked to the verb “show”. This is because it is an operator. An important feature of

RRG from a parsing point of view is that parsing happens in two projections: the constituent projection, shown in figure 1 and the operator projection, which consists of words which modify other words (such as auxiliaries and adjectives). This is important because modifiers are often optional and it simplifies the parsing process considerably if these can be handled separately. Note that adverbs, which can modify larger constituents (such as COREs and CLAUSES) go in the constituent projection so that it is clear what they are modifying. “Yesterday” in this example is an adverb which modifies the CORE, to show when the action took place.

RRG makes extensive use of templates. These templates consist of whole trees and are thus harder to use in a parsing algorithm than rules. The templates can easily be reduced to rules, but only at a loss of much important information. The example in figure 1 consists of one large template that gives the overall structure and some simple templates (which are equivalent to rules) so that elements such as NP and PP can be expanded. An NP is a noun phrase and in this theory consists of a noun, pronoun, or question word. Templates are required to parse complex noun phrases, such as those with embedded clauses. A PP is a prepositional phrase and consists of a preposition followed by a NP. Clearly if we reduce the template in the example in figure 1 to the rule

CLAUSE → NP NP V PP ADV

we lose a lot of the information inherent in the structure of the template. A further feature of RRG is that the branches of the templates do not have to have a fixed order and lines are allowed to cross. The latter is important for languages such as German and Dutch where the adverb that makes up the periphery normally occurs within the core. This feature will be important in our application for marking work by students for whom English is not their first language.

The above features pose challenges for parsing according to the RRG paradigm. We have overcome these challenges by making some additions to the standard chart parsing algorithm. The main innovations are

- a) a modification to enable parsing with templates
- b) a modification to allow variable word order.

In addition, parsing also includes elements of dependency grammar to find operators and to determine which word they belong to. At present the most popular methods of parsing are HPSG (Hou and Cercone, 2001, Kešelj, 2001, Wahlster, 2000) and dependency grammar (Chung and Rim, 2004, Covington, 2003, Holan, 2002). HPSG is good for fixed word order languages and dependency grammar is good for free word order languages. The approach to parsing described below is novel in that it allows parsing with templates, and because of the range in flexibility of word order allowed.

## Outline of the parsing algorithm

The parsing algorithm relies on correctly tagged text. We use Shoebox (available from SIL ([www.sil.org/computing/shoebox](http://www.sil.org/computing/shoebox))) to tag sentences. Shoebox is a semi-interactive tagging program. It was chosen because the user can define their own tags and because it is easy to ensure all tags are correct. This is a good program to use for experimentation. Once the tags have been finalised an appropriate automatic tagger can be used, or written using standard techniques.

Once a sentence has been tagged, there are three parts to the parsing algorithm:

1. **Strip the operators.** This part removes all words that modify other words. It is based on a correct tagging of head and modifying words. This stage uses methods from dependency grammar and the end result is a simplified sentence.
2. **Parse the simplified sentence using templates.** This is done by collapsing the templates to rules, parsing using a chart parser and then rebuilding the trees at the end using a complex manipulation of pointers. The chart parser has been modified to handle varying degrees of word order flexibility. This is done by working out all the possible combinations of the ordering using breadth first search. These options are then built into a complex data structure in such a way that relevant parts are deleted as parsing progresses, leaving the correct option according to the data.
3. **Draw the resulting parse tree.**

Details of the extensions to the chart parser are given below.

## Parsing Templates

Templates are parsed by collapsing all the templates to rules and then rebuilding the correct parse tree once parsing is complete. This is done by including the template tree in the rule, as well as the left and right hand sides. When rules are combined during parsing, we make sure that the right hand side elements of the instantiated rule, as represented in the partial parse tree, point to the leaves of the appropriate rule template tree. This is especially important when the order of the leaves of the template may have been changed. The reference number for the rule that has been applied is also recorded so that it can be found quickly.

Modifying nodes, such as PERIPHERY, cause problems with rebuilding the tree. This is because such nodes can occur anywhere within the template, including at the root and leaf levels. Also, if we are dealing with a sub-rule whose root node in the parse tree has a modifying node, it is not possible to tell whether this is a hang-over from the previous template, or part of the new template. To solve this problem, modifying nodes have flags to say whether

they have been considered or not. There is a potential additional problem with repeated nested rules because if processing is done in the wrong order, the pointers to the rule template tree get messed up. To overcome this problem, each leaf of a template is dealt with before considering sub-rules.

The algorithm for building the tree is:

1. Get the appropriate rule and rule template tree
2. If the rule tree is of depth 1 and has no embedded modifying nodes (that is modifying nodes that point to a node other than the root), then we can simply continue by looking at each of the children in turn, starting at step 1.
3. If the rule tree is of depth greater than 1 or there are embedded modifying nodes, then make the rule template tree point to the appropriate places in the parse tree. This is done using the links made from the parse tree to the rule template tree during parsing. Note that the parse tree will consist of simple rule structures of depth 1 and modifying nodes will show up as children.
4. Clear all the children in the parse tree. This will have the effect of removing any embedded modifying nodes.
5. Copy all the children of the template tree and copy into the appropriate place in the parse tree.
6. If the template has modifying nodes, copy that part of the template tree and insert into the appropriate place in the parse tree.
7. Replace the leaves of the copied template trees with the original leaves. This is possible because the template leaves are pointing to the original leaves (step 3).
8. Consider each leaf in turn, modifying the parse tree as above (start at step 1 for each leaf).

### **Parsing with fixed, free, and constrained word order**

There were two main problems to solve in order to modify the chart parser to handle varying degrees of word order flexibility:

1. Working out a notation for denoting how the word order can be modified.
2. Working out a method of parsing using this notation.

(1) was achieved by the following notation on the ordering of the leaves of the template, treating the template as a rule.

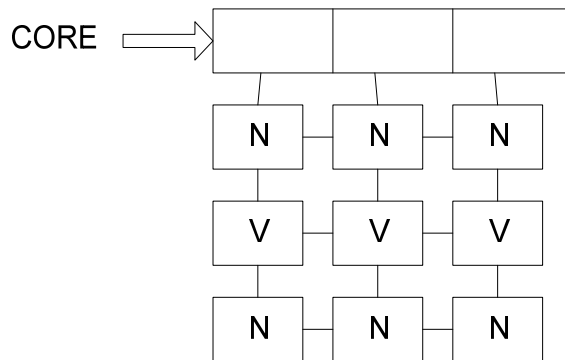
- Fixed word order: leave as it is {N V N}
- Free word order: insert commas between each element {N,V,N} (Note that case information is included as an operator so that the undergoer and actor can be identified once parsing is complete.)
- An element has to appear in a fixed position: use angular brackets: {N, <V>, ADV} this means that N and ADV can occur before or after

v, but that V MUST occur in 2<sup>nd</sup> position. Note that this is 2<sup>nd</sup> position counting constituents, not words.

- Other kinds of variation can be obtained via bracketing. So for example {(N, V) CONJ (N, V)} means that the N's and V's can change order, but that the CONJ must come between each group. If we had {(N,V),CONJ,(N,V)} Then the N's and V's must occur next to each other, but each group doesn't have to be separated by the CONJ, which can occur at the start, in the middle, or at the end, but which cannot break up an {N,V} group.

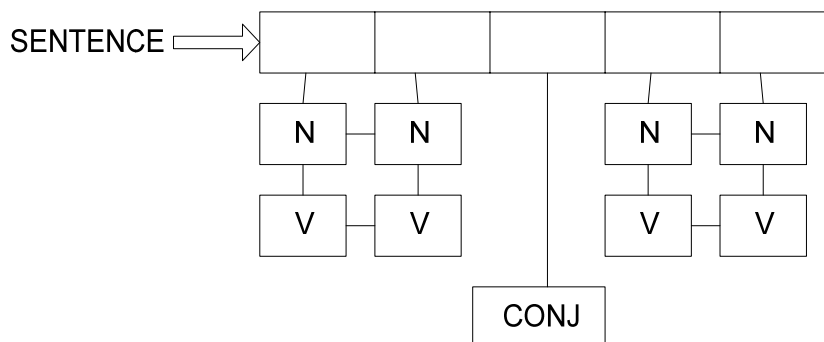
### Modifications to the parsing algorithm.

Parsing was achieved via a structure that encoded all the possible orderings of a rule. So for example the rule CORE→N, V, N would become



This means that N or V can occur in any position and N has to occur twice. The lines between the boxes enable the “rule” to be updated as elements are found.

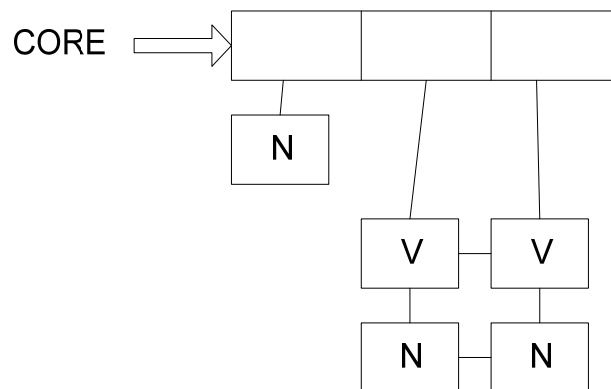
Using this schema, SENTENCE→(N,V) CONJ (N,V) would become



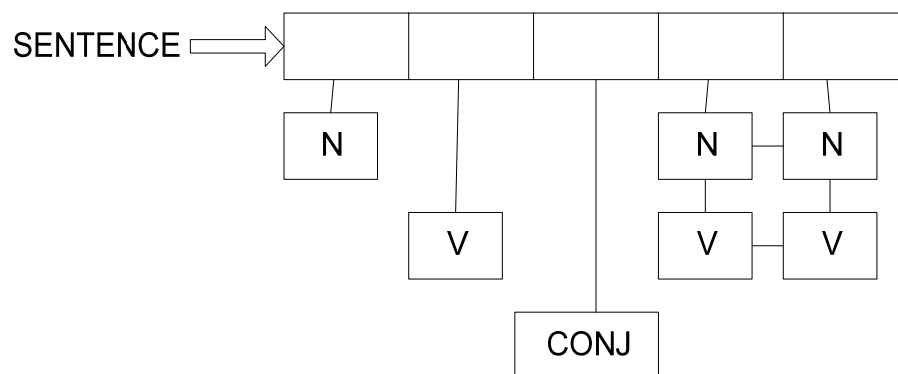
In this case, the CONJ in the middle is by itself because it has to occur in this position because the grouping word order is fixed. The groupings of N's and V's show where the free word ordering can occur.

To apply a rule, the first column of the left hand side of the rule is searched for the token. Any tokens that do not match are deleted along with the path that

leads from them. In the first example, after an N is found, we would be left with



And in the second example, after an N is found we would be left with



Note that in order for the rule to be satisfied, we *must* find a V and then a CONJ: there are no options for position 2 once the element for position 1 has been established.

In this way, we can keep track of which elements of a rule have been found and which are still to be found. Changes in ordering with respect to the template are catered for by making sure that all instantiated rules point back to the appropriate leaves of the rule template, as described above.

The different possibilities for each rule are obtained via a breadth first search method that treats tokens in brackets as blocks. Then the problem becomes one of working out the number of ways that blocks of different sizes will fit into the number of slots in the rule.

## Results

Preliminary results of applying these algorithms to student texts are very promising, but some issues have been highlighted. The method parses relatively simple sentences correctly and the main arguments and verbs are

found. In addition, some very long and complicated sentences are parsed correctly and many kinds of grammatical errors do not cause any problems.

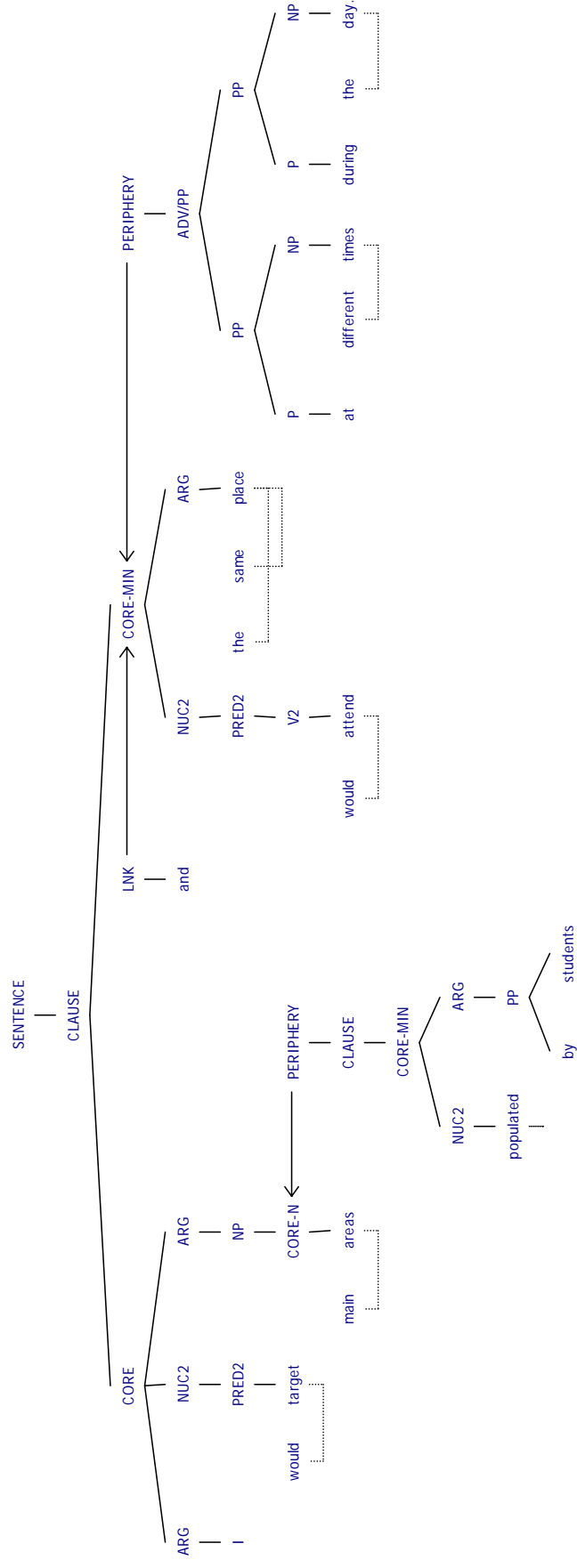


Figure 2: An example of a correctly parsed sentence.



An example of a correctly parsed sentence is “I would target main areas populated by students and would attend the same place at different times and during the day.” The parse tree for this example is given in figure 2. Note that the complex object “main areas populated by students” has been parsed correctly and that the tree attaches the qualifying phrase to “area” so that it is clear what is being qualified. An important source of ambiguity in English sentences is caused by prepositional phrases and this is a main cause of multiple parses of a sentence. In this example, the phrases “at different times” and “during the day” are placed together in the periphery of the CORE, although arguably they should have a different structure. This is a design decision to limit the number of parses. This kind of information needs semantic information to sort out what attaches to what. This cannot be obtained purely from the syntax.

An example of an ungrammatical sentence that is correctly parsed is “Results from the observations would be less bias if the sample again was not limit the students in the labs between 9:30 and 10:30 on a Thursday morning.” This sentence parses correctly because the affix that should be on “limit” is an operator and the correctness of the operators is not checked during the parsing process. The word “bias” is labelled as a noun and gets attached as the second argument to “would be”, although it should be “biased”, which would get it labelled as an adjective. Despite these errors, the meaning of the sentence is clear and the parse will enable the meaning to be deduced.

The sentence “Therefore, asking only the students present on a Thursday morning will exclude all the students that either have no lessons or are not present” produces two parses: once correct and one incorrect. The incorrect parse breaks up “Thursday morning” to give two clauses:

- a. Asking only students present on a Thursday
- b. Morning will exclude all the students that either have no lessons or are not present

In the first clause, the subject is “asking only students”, the main verb is “present” and the object is “on a Thursday morning”. This does not make sense, but it is syntactically correct as far as the main constituents are concerned. Similarly, the second clause is also syntactically correct, although it does not make sense. There are two ways of eliminating this parse. The first is to do a semantic analysis; the second is to not allow two clauses juxtaposed next to each other without punctuation such as a comma. However, students tend to not be very good at getting their punctuation correct. The current implementation of the parsing algorithm ignores all punctuation other than full stops for this reason.

An issue that makes parsing problematic is that of adverbs. These tend to be allowed to occur within several places within the core and some, such as yesterday, modify groups of words rather than a single word. The best solution, given their relative freedom of placing and the fact that sorting out where best to put them is more a meaning than a syntactic issue, would be to

remove them and work out where they belong once the main verb and arguments have been identified.

Most of the above issues have to be left to an analysis of meaning to sort out the correct parse. There is no clear division between syntax and semantics. However there is another issue that has been highlighted to do with grammar and punctuation. How tolerant of errors should the system be? We have shown that errors in the operators do not cause problems for the parser, and errors in the placing of adverbs are relatively easy to deal with, but errors in the main constituents are not handled. For example the phrase “the main people you need to ask will not be in the labs so early unless that have got work to hand in” occurs in one of the texts. The current algorithm will not handle these kinds of mistakes. But should the system be able to handle these kinds of mistakes, or should students be encouraged to improve their writing skills?

## **Conclusion**

We argue that this approach, though still under development, potentially has huge benefits for students and staff in higher education and could, with further improvements, form one building block in constructing a new paradigm for CAA. Our intention is to use this as the first stage in a system that uses a new semantic framework, ULM (Universal Lexical Metalanguage) (Guest and Mairal Usón, 2005), to compare the meaning of student texts with a (single) model answer.

## References

- Beevers, C E, Foster, M G, and McGuire, GR. 1989. Integrating Formative Evaluation into a Learner Centred Revision Course. *British Journal of Educational Technology*:115-119.
- Brown, S, Race, P, and Bull, J. 1999. *Computer Assisted Learning in Higher Education*. London: Kogan Page.
- Chung, Hoojung, and Rim, Hae-Chang. 2004. Unlexicalized Dependency Parser for Variable Word Order Languages based on Local Contextual Pattern. [Feb 15-21]. *Lecture Notes in Computer Science: Computational Linguistics and Intelligent Text Processing (5th International Conference CICLING)* 2945:112-123.
- Covington, Michael A. 2003. A Free Word Order Dependency Parser in Prolog.
- Guest, E, and Mairal Usón, Ricardo. 2005. Lexical Representation Based on a Universal Metalanguage. *RAEL, Revista Española de Lingüística Aplicada* 4:125-173.
- Holan, Tomáš. 2002. Dependency Analyser Configurable by Measures. *Text, Speech and Dialogue 5th International Conference TSD*:81-88.
- Hou, Lijun, and Cercone, Nick. 2001. Extracting Meaningful Semantic Information with EMATISE: an HPSG-Based Internet Search Engine Parser. *IEEE International Conference on Systems, Man, and Cybernetics* 5:2858-2866.
- Kešelj, Valdo. 2001. Modular HPSG. *IEEE International Conference on Systems, Man, and Cybernetics* 5:2867-2872.
- Landauer, Thomas K., Laham, Darrell, Rehder, Bob, and Schreiner, M. E. 1997. How well can Passage Meaning be Derived without using Word Order? A Comparison of Latent Semantic Analysis and Humans. *Proceedings of 19th Annual Conference of the Cognitive Science Society*:412-417.
- Pérez, D, and Alfonsa, E. 2005. Adapting the Automatic Assessment of Free-Text Answers to the Students. Paper presented at *9th Computer Assisted Assessment Conference*, Loughborough, UK.
- Rust, C. 2002. The Impact of Assessment on Student Learning. *Active Learning in Higher Education* 3:145-158.
- Sadler, D R. 1989. Formative Assessment and the Design of Instructional Systems. *Instructional Science* 18:119-144.
- Sadler, D R. 1998. Formative Assessment: Revisiting the Territory. *Assessment in Education: Principles, Policy and Practice* 5.
- Sambell, K, and Hubbard, A. 2004. The Role of Formative 'Low Stakes' Assessment in Supporting Non-Traditional Students' Retention and Progression in Higher Education: Student Perspectives. *Widening Participation and Lifelong Learning* 6:25-36.
- Sukkarieh, Jana Z, Pulman, Stephen G, and Raikes, Nicholas. 2003. Auto-marking: using computational linguistics to score short, free text responses. Paper presented at *International Association of Educational Assessment*, Manchester, UK.
- Sukkarieh, Jana Z, Pulman, Stephen G, and Raikes, Nicholas. 2004. Auto-Marking 2: An Update on the UCLES-Oxford University research into

- using Computational Linguistics to Score Short, Free Text Responses.  
Paper presented at *International Association of Educational Assessment*, Philadelphia.
- Van Valin, Robert D Jr, and LaPolla, R. 1997. *Syntax: Structure, Meaning and Function*. Cambridge: Cambridge University Press.
- Van Valin, Robert D Jr. 2005. *Exploring the Syntax-Semantics Interface*: Cambridge University Press.
- Wahlster, Wolfgang. 2000. *Verbmobil: Foundations of Speech-to-Speech Translation*: Springer.
- Wiemer-Hastings, Peter. 2001. Rules for Syntax, Vectors for Semantics. *Proceedings of 22nd Annual Conference of the Cognitive Science Society*.
- Yorke, M. 2001. Formative Assessment and its Relevance to Retention. *Higher Education Research and Development* 20:115-126.

# **FROM ONLINE ENTRIES TO ONLINE RESULTS**

**(DEVELOPING AN INTEGRATED E-  
ASSESSMENT SYSTEM LINKING  
INTERNET DELIVERY OF A TEST  
WITH BACK-END ELECTRONIC  
PROCESSING SYSTEMS)**

**Ed Hackett and Paul Seddon**



# **From Online Entries to Online Results**

Ed Hackett – Examinations Manager, Assessment and Operations Group; Paul Seddon – CBT Programme Manager, Customer Services Group, University of Cambridge ESOL Examinations (UCLES)

## **Introduction and background**

In November 2005, University of Cambridge ESOL Examinations (Cambridge ESOL) launched an internet delivered computer-based version of the Preliminary English Test (PET). Since then, a number of wraparound packages have been introduced to enable centres to make entries and receive results online. In autumn 2007, with the introduction of on screen marking, the final piece of the e-assessment jigsaw will be put in place, providing Cambridge ESOL and its centres with the complete integrated e-assessment package. Further products have now been added to this online delivery system, including tests from other Cambridge Assessment business streams, OCR and CIE (University of Cambridge International Examinations). This paper outlines some of the key development stages undertaken and discusses a number of issues arising out of these developments, both in terms of the questions they raised and the action subsequently taken. It also explores issues that merit further discussion, research or development.

Cambridge ESOL has produced computer-based tests since 2000, but prior to the launch of CB PET in November 2005, these were all CD-ROM based. PET is a general English examination for speakers of other languages and is at level B1 in the Council of Europe framework of reference and Entry Level 3 in the UK National Qualifications Framework. It tests four skills: reading, writing, listening and speaking. Paper-based (PB) PET was introduced in the late 1970s and was most recently updated in format in 2004. With a fast growing candidature, a 45% increase since 2000, and a young exam population, over 70% of candidates aged under 20, it was felt that PET was an appropriate choice of exam for conversion to a computer-based product.

## **Developing and integrated e-assessment system**

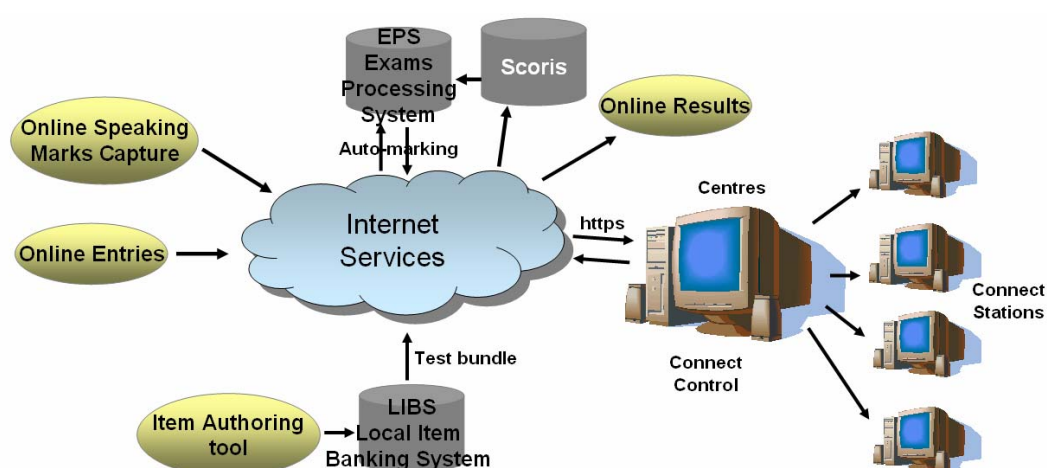
With the vast majority of Cambridge Assessment's examinations being paper-based, it was important to develop a system which could integrate with existing exams processing systems. This inevitably raises issues with legacy systems. Do you try to enhance the capabilities of the existing system or is it better to bypass it and develop additional software to meet all the necessary

requirements? Often, there is no choice, but to adapt the existing systems, and this can prove both problematic and costly. Furthermore, the issue of IT resource also has to be factored in. Do you wait until there is sufficient resource and budget for every part of the jigsaw to be put into place, or do you develop the product piecemeal, developing the key functional elements first and bringing forward the launch date?

## Technical Developments

Cambridge ESOL developed its generic online delivery engine, Cambridge Connect, in a phased approach; the primary phase being customer/candidate centric enabling the delivery of a test to candidates over a distributed network. The over-arching requirement was for a delivery engine specifically purposed for the delivery of high-stakes examinations worldwide (i.e. internationally recognised exams with a high surrender value that can be used for immigration purposes or school leaving certification for example). As such, there could be no opportunities for a test to be affected by variations in internet connectivity which therefore dictated that whilst the exam could be delivered online, it was downloaded prior to the examination and taken offline.

Cambridge Connect is primarily focussed on test delivery and as such is customer facing; but this is only half the story. Cambridge Connect needs to integrate with back end processing systems such as our Local Item banking System (LIBS) and the Exams Processing System (EPS), which handles candidate entries, marks capture and the processing of results. In addition, Connect integrates with numerous other systems to enable marking and processing from end-to-end in order to create a seamless paperless experience for Centres and Candidates.



**Figure 1. The Connect Framework**



## How does it all work?

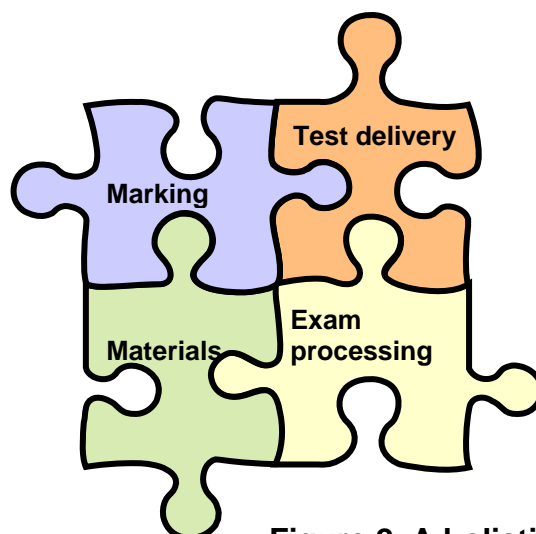
Within the item bank, pre-tested items are copied into the Item Conversion Tool (ICT). This tool marks up the items in QTi XML, enabling them to be read by the Connect delivery engine, and publishes an electronic test bundle to the Connect hub, a series of web services customers don't see.

For the centres, the experience starts with making Entries, which are keyed in online. Entries are linked to session data in the Exams Processing System and are then communicated along with eligible centre details to the Connect hub.

At the Centre, the Connect software is installed on a network and, at a pre-defined time before the start of the test date, centres can download an encrypted test bundle via https protocols. This test stays encrypted until the test is ready to start on the test day; candidates are provided with login details printed from Connect and start the test. Connect has a number of failsafe features built-in in the event of computer failure. If a candidate's PC fails then the candidate can simply be moved to another PC and resume where they left off. If the Connect Control PC (the PC on which the exam management software runs) fails, a backup recovery tool enables the test administrator to resume the test.

At the end of the test, the candidates' responses are encrypted and uploaded directly to Cambridge web servers at our Data Centre, where different marking applications are employed depending on the type of exam or item types. Some exams consisting of multiple choice question types and short answer responses can be fully automarked; others use the on screen marking application (Scoris), part of Electronic Script Marking system (ESM), enabling examiners to call up candidates' written responses and mark them on screen. Marks are then aggregated and returned into EPS for scaling, grading and results and certificate production.

Cambridge Connect therefore introduces a new and holistic approach (figure 2) to the production, delivery and processing of Cambridge Assessment exams.



**Figure 2. A holistic approach to e-assessment**

## Test Development and Construction

Converting an existing paper-based test for on-screen delivery is very different to the process of developing a computer-based test from scratch. In the latter, you have free rein to develop and trial tasks that you feel best fit this medium. In the former, you have to decide whether the computer-based test is going to follow the same format as the paper-based version and to what extent both modes will be comparable in terms of reliability and results. It was important to Cambridge ESOL that the computer-based test variant of the exam did not advantage or disadvantage candidates when compared to the PB format, and that a grade obtained via the CB mode would have the same value as the equivalent grade obtained using the traditional PB method. A decision was also made to retain the face-to-face format of the Speaking test, though the introduction of online marks capture would allow examiner marks to be keyed into a web application and returned electronically. The key aim was then to prove that it would be possible to transfer the format and task types used in the PB Reading, Writing and Listening tests to an on screen variant.

Four key stages of development were identified:

- feasibility study;
- task design and trialling;
- navigation design and trialling;
- equivalence trialling.

The aim of the feasibility study was to look at the suitability of the tasks in the Reading and Writing and Listening components for on-screen adaptation and to propose designs for trialling. Cambridge ESOL has produced computer-based tests in CD-ROM format since 2000, for example CB BULATS (Business Language Testing Service) and QPT (the Quick Placement Test, which is marketed by Oxford University Press), and development work had already been done on CB IELTS (International English Language Testing System) -launched in May 2005, so a certain amount of knowledge and expertise had already been gained from the development and use of these products.

One of the key issues in converting paper-based materials for on-screen delivery is the use of the computer screen real estate. For example, in a paper-based test the candidate can view two pages of text at one time, whereas a computer screen can only display part of this text at any one time. In addition to this, part of the screen in a CB test is taken up with navigation buttons. This does not present a problem for discrete tasks, tasks with only one item, which can be displayed on screen in their entirety, e.g. PET Reading Part 1 and PET Listening Part 1 (see *table 1 below*), where the task consists of one or more small graphics, one short question and 3 multiple choice options. However, in grouped-question tasks, decisions had to be made over the display of longer text and question input.

**Table 1: CB PET Test Content for Reading, Writing and Listening**

<b>READING</b>			
<b>Part</b>	<b>Task Type and Format</b>	<b>Task Focus</b>	<b>Marking Method</b>
<b>1</b>	Three-option Multiple choice discrete. <i>Five</i> very short discrete texts: signs and messages, postcards, notes, e-mails, labels etc., plus one example.	Reading real-world notices for main message.	Automarked
<b>2</b>	Matching – grouped task <i>Five</i> items in the form of descriptions of people to match to eight short authentic-adapted texts.	Reading multiple texts for specific information and detailed comprehension	Automarked
<b>3</b>	True/False – grouped task <i>Ten</i> items with an adapted-authentic long text.	Processing a factual text. Scanning for specific information while disregarding redundant material.	Automarked
<b>4</b>	Four-option multiple choice – grouped task. <i>Five</i> items with an adapted-authentic long text.	Reading for detailed comprehension; understanding attitude, opinion and writer purpose. Reading for gist, inference and global meaning.	Automarked
<b>5</b>	Four-option Multiple-choice – grouped task. <i>Ten</i> items, plus an integrated example, with an adapted-authentic text drawn from a variety of sources. The text is of a factual or narrative nature.	Understanding of vocabulary and grammar in a short text. Reading for general and detailed meaning, and understanding the lexico-structural patterns in the text.	Automarked
<b>WRITING</b>			
<b>Part</b>	<b>Task Type and Format</b>	<b>Task Focus</b>	<b>Marking Method</b>
<b>1</b>	Sentence transformations. <i>Five</i> items, plus an integrated example, that are theme-related. Candidates are given sentences and then asked to complete similar sentences using a different structural pattern so that the sentence still has the same meaning.	Control and understanding of Threshold/PET grammatical structures. Rephrasing and reformulating information.	Automarked

<b>2</b>	Short communicative message. Candidates are prompted to write a short message in the form of a postcard, note, e-mail etc. The prompt takes the form of a rubric or short input text to respond to.	A short piece of writing of 35 - 45 words focusing on communication of specific messages.	On Screen marking
<b>3</b>	A longer piece of continuous writing. Candidates are presented with a choice of two questions, an informal letter or a story. Candidates are primarily assessed on their ability to use and control a range of Threshold-level language. Coherent organisation, spelling and punctuation are also assessed.	Writing about 100 words focusing on control and range of language.	On Screen marking
<b>LISTENING</b>			
<b>Part</b>	<b>Task Type and Focus</b>	<b>Task Format</b>	<b>Marking Method</b>
<b>1</b>	Multiple choice (discrete). Short neutral or informal monologues or dialogues. <i>Seven</i> discrete three-option multiple choice items with visuals, plus one example.	Listening to identify key information from short exchanges.	Automarked
<b>2</b>	Multiple choice – grouped task Longer monologue or interview (with one main speaker).  <i>Six</i> three-option multiple choice items.	Listening to identify specific information and detailed meaning.	Automarked
<b>3</b>	Gap-fill – grouped task Listening to identify, understand and interpret information. Using this information to fill <i>six</i> gaps on a form or to complete notes.	Longer monologue of neutral or informal nature.	Onscreen Marking
<b>4</b>	True/false – grouped task Longer informal dialogue. Candidates need to decide whether <i>six</i> statements are correct or incorrect.	Listening for detailed meaning, and to identify the attitudes and opinions of the speakers.	Automarked

Decisions over the use of pagination, used in the older CD-ROM format tests, and scrolling, the most common format for websites, had to be made. The colour and size of font and background screen colour were also important factors, as was the format of the graphics. Furthermore, onscreen rendering

of the tasks had to be integrated with items drawn from the current paper-based item bank, which meant converting word-based tasks into XML.

The feasibility study revealed that it should be possible to represent all the paper-based tasks on screen and task, navigation and equivalence trialling revealed few major problems. As anticipated, an overall preference for taking PET on computer was expressed by the majority of candidates taking part in equivalence trialling (190 candidates in 4 different countries). 63% preferred taking the Reading and Writing test on computer, as opposed to 20% preferring the paper-based version. For the Listening test, 83% expressed a preference for the computer version, with only 4% preferring the paper test (Hackett, 2005). Candidates found the proposed functionality for answering both multiple choice and typed answers clear and easy to use. Following task trialling, the additional functionality to remove a multiple choice answer already entered was added. This allows candidates to leave a question unanswered, having already entered an answer, should they want to leave it blank and return to it later. It was also discovered that some candidates at this level had difficulty following a grouped listening task and typing answers at the same time (PET Listening Part 3). Candidates were subsequently allowed to make notes on paper and were given additional time to type these up at the end of the task.

No major problems were identified with reading text on screen, though a number of candidates did express a desire to be able to highlight text. This has been backed up by feedback from some candidates taking the test in live sittings, though no drop in reading scores on the CB mode has been identified. Cambridge Assessment is investigating the technology necessary to add this functionality for a future release of Connect. Further research into the impact of reading on screen versus reading on paper is high on the agenda at Cambridge ESOL. In response to the question, 'Did you find reading on computer easier than reading on paper?', 46% found it easier, whereas only 25% preferred reading on paper. This perhaps reflects an increasing familiarity with on-screen reading, at home, in school or at work. PET, as a level B1 test, has a limited reading load for candidates, with the maximum length of text being 450 words. Higher level exams with longer reading passages will exert greater strain on the reader and might impact on the task. Paek (2005), in reviewing CB and PB versions of tests in the American schools sector, noted that extended reading passages tended to appear more difficult in CB format. This is clearly an area warranting further research and the introduction of new examinations to the Connect delivery system will help provide more data for analysis.

Writing also proved more popular on screen, with 67% showing a preference for typing and only 25% expressing a preference for handwriting. CB PET disables grammar and spell checks in an effort to maintain the conditions of the PB equivalent, though the screen does include a word count. However, if we were to attempt to replicate real-life writing situations, it could be argued that grammar and spell check facilities ought to be included. This would necessitate the introduction of a separate markscheme reflecting the resulting improved standards of accuracy and may cause problems in differentiating

between candidates who are naturally able to use language accurately and those who are able to exploit the correction aids available. For the Writing section, other key issues were the impact of typing on candidate performance, and the affect of type-written script on examiner marking; i.e. do examiners treat typed script more harshly or leniently than handwritten script? A number of studies into this area have been carried out for CB IELTS (Thighe et al, 2001, and Green and Maycock, 2004), but given the different test format and candidature, it was agreed that further validation studies would need to be carried out. The benefits of using new marking procedures and analytical tools made available by the advent of on screen marking are explored further in section 5.

## **Marking and Grading**

As mentioned above, development of a fully integrated system was split into various phases, with online test delivery preceding electronic marking of responses. The traditional method for marking Cambridge PB tests is via an optical mark reader (OMR) answer sheet. The candidate lozenges in multiple choice answers and writes any written responses within defined spaces on the answer sheet. On return to Cambridge ESOL, written responses are marked either by general markers e.g. for short responses, or by examiners, for longer composition type answers e.g. the letter or story in PET Writing Part 3. The general marker or examiner lozenges a score on the OMR, which is then scanned into the exams processing system, where multiple choice answers are electronically auto-marked against a pre-populated key and added to general and examiner marks. Speaking marks are entered by the examiner onto an OMR and this is returned to Cambridge for scanning. Item level data can then be extracted by the Validation department ahead of grading.

In phase 1 of the project, candidate responses were overprinted onto OMRs so that written responses could be marked in the same way as PB responses, with the OMRs then being scanned. Speaking marks were collected in the same way as for PB (above). The development of an online portal for entering speaking marks at source, in November 2006, meant that speaking marks could be directly ported to the exams database. The introduction of this facility negated the need to print and despatch speaking OMRs to centres prior to the exam and the need for centres to return these marksheets to Cambridge, speeding up the back-end processing of scores and reducing the entry window by 2 weeks.

The final phase of development is the introduction of on screen marking for human rated tasks. This not only allows the speeding up of the marking process, but offers the opportunity for improvements to the examiner marking system, developing online support for markers and contributing to increased rating reliability. In parallel with this system, multiple choice and some short answers will be directly automarked, without the need to print to OMR and scan. The other short productive items, those deemed too complex to be

automarked, will be delivered onscreen to general markers. Longer texts will be routed to examiners, who will undergo co-ordination and standardisation and mark via their home computers. On screen marking has already been developed and used by both OCR and CIE for marking PB products, where completed exams papers are first scanned. For CB, there are obvious cost savings, as responses do not require scanning. The responses returned via Connect are displayed to the examiner using same screen view that the candidate sees.

However, one of the additional advantages of using on screen marking is not simply savings in time or cost. It is the opportunities it offers for the implementation of new examiner marking models, that is particularly interesting. There are various models employed for examiner marking of Cambridge ESOL exams, utilising both on-site and at-home marking scenarios. PB PET is currently marked on-site using a partial remarking model. Examiners are put into teams which are monitored by a team leader, who in turn reports to a Principal Examiner. Following co-ordination and standardisation, each examiner is monitored by the team leader, who informs the examiner of leniency, harshness or erratic performance early on in the process. The aim of this approach is that performance is monitored and modified where necessary. Batches of scripts are then remarked where appropriate and monitoring continues over the marking weekend. At home marking models also include co-ordination and standardisation, in addition to batch sampling. Examiner marks are then subject to scaling, to take account of identifiable leniency or harshness. A third model is double marking, with both examiner marks being averaged, or those deemed outside acceptable tolerance, i.e. differing by too great a margin, being sent to a third, experienced, rater.

On screen marking offers the opportunity for a new model of marking and the possibility of greater intervention in examiner marking behaviour. In addition to the use of co-ordination and standardisation scripts as processes designed to appropriately align examiner behaviour, there is also the possibility of using seeded 'gold standard' scripts (Shaw, 2007) as a means of monitoring such behaviour. Gold standard scripts are candidate samples specially selected as models for use in blind monitoring. These scripts are selected and pre-marked by the PE and a group of senior team leaders, and then seeded as ordinary unmarked scripts into the marking pool each examiner gets. The Principal Examiner or Team Leader is then able to monitor, at various stages during the marking, the relationship between the agreed marks for these scripts and those given by different examiners, and feed this information back into the marking process as a means to achieving greater reliability between markers. Furthermore, the electronic capture of interim as well as final marks provides the validation group with valuable information that can feed into future research. Shaw (2007) identifies a number of interesting research questions that would benefit from the capture of this data:

- In what ways do raters differ? Is there a gender effect? (Facets of Rater Status, Rater Profile and Rater Behaviour.)

- Is it possible to identify distinct rater types and certain patterns of rater behaviour? (Facet of Rating behaviour.)
- What amount of training/re-training is required? Can training improve raters' self-consistency? (Facets of Rater Behaviour and Rater Training.)
- How does assessment differ when marking electronically as opposed to paper-based marking? (Facets of Rater Profile and Rater Behaviour.)

Shaw goes on to state that the data gathered from such exercises could also be used to establish whether particular raters favoured candidates from a particular L1 background or could be used to investigate further the relationship between the tasks, the candidates and examiners. In PET Writing Part 3, candidates are given a choice between writing a letter or a story. We can now investigate further the question of whether rater reliability varies according to the task, and if certain examiners have greater reliability marking one task type as opposed to another. It may then be possible to allocate certain task types to particular types of raters.

On screen marking for CB products using the Connect delivery engine is scheduled for autumn 2007, so we are unable to comment on the live implementation of this software. On screen marking will, however, provide the final link in our online delivery and processing system, leading to a fully integrated e-assessment package.

## **Future development and research**

Computer-based assessment using a system like Cambridge Connect raises a multitude of research opportunities that are likely to impact on the way we assess candidates in the future.

With computer-based assessment we have a clear insight into the examination process from a candidates' point of view that until now has been impenetrable. We can log each and every key stroke a candidate makes and are able to determine:

- which questions a candidate attempted first
- which questions a candidate returned to, changed their answers etc
- how long a candidate spent on each question
- whether two candidates sitting next to each other input the same answers at the same time

### *6.1 Where might this take us?*

Cambridge Connect, together with the on screen marking application, will provide a wealth of information for formative assessment, for building diagnostic assessments and providing scaffolding to help the candidate. It could enable assessment organisations to measure candidates' abilities not



just by getting the answer correct, but also on how long it took the candidate to come up with the correct answer and therefore award additional marks for speediness. It enables assessment organisations to pinpoint a candidate's ability or knowledge by tracking which tasks candidates struggled with or conversely, which tasks are not measuring or performing well in a test because a whole cohort struggled with it. Furthermore, the possibilities for live item calibration (live pre-testing) by seeding uncalibrated tasks into a live exam offered by computer-based assessment enables both exam boards and candidates to reap the benefits and achieve even more meaningful measurement of candidates and their abilities.

## **Conclusion**

In developing Cambridge Connect and integrating it with both existing processing systems and newly developed wraparound e-services, Cambridge Assessment can now deliver high stakes examinations worldwide, achieving vastly reduced entry and results processing times. We are also in the position to explore more fully the comparability of computer-based tests with their traditional paper-based equivalents, and how the differing modes impact on both candidate and examiner behaviour and performance. As Jones (2007) states, 'It is important for Cambridge ESOL to define an approach to comparability which will guide the validation of ...(new CB examinations using Cambridge Connect), ... while providing a more general framework for thinking about comparability of technology-based and traditional assessment.' It therefore hoped that a greater understanding of the candidate experience, in terms of their interaction with computer-based tests, will inform the development of future computer-based tasks and tests.

## References

Green, A and Maycock, L (2004) *Computer-based IELTS and paper-based IELTS*, Research Notes 18, November 2004 (UCLES).

Hackett, E (2005) *The Development of a Computer-based version of PET*, Research Notes 22, November 2005 (UCLES).

Jones, N (2007) *The comparability of computer-based and paper-based tests: goals, approaches, and a review of research*, Research Notes 27, February 2007 (UCLES)

Paek, P (2005) *Recent Trends in Comparability Studies*, PEM Research Report 05-05, [Pearsonedmeasurement.com/research/research.htm](http://Pearsonedmeasurement.com/research/research.htm)

Seddon, P (2005), *An overview of Computer-based PET*, Research Notes 22, November 2005 (UCLES)

Shaw, S.D (2007) *Modelling facets of the assessment of Writing within an ESM environment*, Research Notes 27, February 2007 (UCLES).

**AN IMPROVED COMPUTER-  
ASSISTED TEST FOR ACCESSIBLE  
COMPUTER-ASSISTED  
ASSESSMENT**

**Gill Harrison and John Gray**



# **An Improved Computer-Assisted Test for Accessible Computer-Assisted Assessment**

Gill Harrison  
Innovation North  
Leeds Metropolitan University  
Headingley Campus  
Leeds  
LS6 3QS  
g.harrison@leedsmet.ac.uk

John Gray  
Innovation North  
Leeds Metropolitan University  
Headingley Campus  
Leeds  
LS6 3QS  
j.gray@leedsmet.ac.uk

## **Abstract**

This paper builds on work carried out in the development of a computer-assisted test to be used for staff development purposes (Harrison and Gray, 2006). The test is designed to raise staff awareness of disability issues in relation to the use of technology and of CAA, and includes attempts to simulate some of the experiences of disabled people. Some staff groups have now experienced the test, and it has been improved in the light of feedback.

## **Introduction**

A computer-assisted test for staff development purposes has been developed (see Harrison and Gray, 2006) and subjected to initial trials at Leeds Metropolitan University. This has been done under the auspices of the Centre for Excellence in Teaching and Learning – Active Learning in Computing, known as CETL ALiC (Durham University, 2006 and Leeds Metropolitan University, 2007). The test is designed to raise staff awareness of disability issues as they relate to the use of technology and CAA, though its aims are relatively modest in comparison with those of some staff development initiatives in other universities (see for example Pearson and Koppi, 2006). It presents questions and gives appropriate feedback on answers. Some

evaluation of the test has now taken place, improvements have been made and plans for the future have been formulated.

### **First version and first run of test**

In June 2006, the test was tried out for the first time on a group of 15 staff from two faculties of Leeds Metropolitan University, Innovation North (the Faculty of Information and Technology) and the Carnegie Faculty of Sport and Education. A session of two hours was used, with the test plus associated discussion occupying the first hour, and the completing of evaluation forms together with lunch occupying the second. (This ensured a very high response to the evaluation!) The session took place in a room with fixed PCs, and was run by three members of the CETL ALiC team, John Gray, Gill Harrison and Jakki Sheridan-Ross (the Research Officer). An introductory talk explaining the aims and format of the session was given, and then the practical part of the session was launched. The questions in the test were designed to try to simulate the experiences of disabled people, for example by showing how a question might appear to a person with a visual impairment. Questions in the test related to motor and cognitive impairments (especially dyslexia) as well as to visual impairments. An option was generally provided to view a question with and without the simulated impairment. See figure 1 below for the test entry page, and figure 2 for a typical question.

Participants were asked to complete each section of the test, visiting the suggested informative web links if they wished, and then to join in a discussion about that section. The interface was very simple, and the number of questions was only 9 (see Figure 1: the questions are shown as underlined, the section headings without underlining). This limited form of the test resulted from some software development difficulties. A decision had been taken not to use the VLE (WebCT) or proprietary CAA software, so as to retain complete freedom in how the test was presented.

In practice, it proved difficult to restrain participants from going through the whole test, once they had started, so the planned structure of the session was revised into a more informal one, with the three presenters talking to small groups of participants as they worked through different parts of the test, and a final plenary discussion.

**Example Questions.**

**Language.**

[Complex Language.](#)

[Letter Reversal](#)

[Similar Words](#)

[Text Decoration](#)

**Visual**

[Tunnel Vision](#)

[Colour Blindness](#)

**Motor Impairment.**

[Moving Items.](#)

[Accurate Positioning.](#)

[Tab Control](#)

Hearing: to be added.

Figure 1: the entry page for the initial test

Colour Blindness	
<p><b>Question Body</b></p> <p>The most common form of colour blindness is between which pairs of colours?</p> <div style="border: 1px solid #ccc; padding: 5px; margin: 10px 0;"> <input type="radio"/> RED / GREEN         </div> <div style="border: 1px solid #ccc; padding: 5px; margin: 5px 0;"> <input type="radio"/> YELLOW / BLUE         </div> <div style="border: 1px solid #ccc; padding: 5px; margin: 5px 0;"> <input type="radio"/> PINK / PURPLE         </div> <div style="text-align: right;"> <input type="button" value="Submit"/> </div>	<p><b>Instructions</b></p> <p>If this is difficult to see, click here  <input type="button" value="Normal View"/> <input type="button" value="Black and White"/></p> <p><b>Comment.</b></p> <p>The overwhelming majority of people who are colour-blind can see colours. They just have difficulty distinguishing between certain colours. Colour blindness is often typified by difficulty in distinguishing between certain colours. Red/Green colour blindness is the commonest form, and more rarely Blue/Yellow colour blindness.</p> <p><b>Guidance</b></p> <p>Be cautious in your use of colour, and allow for colour-blind students.</p>
<p>Useful Links <a href="#">Colour Blindness</a> <a href="#">Colour Contrasts</a></p>	

Figure 2: format of a question in first test

### Evaluation and feedback from first test

A questionnaire, filled in during the hour following the test, was used to elicit feedback. The questions are shown in Figure 3.

Question number	Question
1	Please say whether you feel that the session has increased your awareness of disability issues in relation to technology in general or in relation to computer-assisted assessment (CAA) in particular: (a) increased my awareness of disability/technology in general Yes/ No If yes, in what way(s)? (b) increased my awareness of disability/CAA? Yes/ No If yes, in what way(s)?
2	Regarding the number of questions provided in the test, which of the following would you agree with: (a) too many questions overall (b) about the right number of questions (c) there could be more questions
3	What did you think of the overall session length of one hour? (a) too long (b) about right (c) could be longer
4	What did you most like about the test?
5	What did you most dislike about the test?
6	Do you think that this session will affect your actions in the future? Yes/ No If yes, how?
7	Please state any suggestions for improvements to the session, including additional features or ideas for new questions that you think could be included
8.	Would you recommend this session to others? Yes/No
9.	Would you be interested in attending a more in-depth session about disabilities and Computer Assisted Assessment?
10.	Please use this space for any other feedback you would like to give.

**Figure 3: questionnaire**



13 responses were received (two people had to leave before the evaluation). Responses were generally positive and are discussed in detail below.

*Q1: whether the session had increased the participant's awareness of disability issues in relation to technology and/or CAA, and if so how*

Apart from the Disability Support expert from Learning Information Services, who responded that this was her job so she knew a lot about it already (though not necessarily the assessment side of things), all responded yes to both the technology and the CAA parts of this question.

With regard to technology issues, responses focused on the usefulness of the empathetic aspect of the test: "useful ... to present how a disabled student may feel by putting staff in the position of students". The other main point raised was that people felt they had been reminded of the wide range of impairments that exist, when they had perhaps before only tended to think of a limited number: "learned about different forms/types of dyslexia", "helped me to think of the different types of disability", "new awareness of some of the specifics of various impairments, eg colour blindness".

Regarding specifically CAA issues, responses were more limited. They included "made me realise the extent to which visual and motor impairments can limit performance on certain types of assessment" and "useful to explore CAA and paper based assessment especially in relation to students with text reading problems". One respondent commented "in terms of CAA it showed how important it is to respond to the student straight away with an explanation, good formative style", which does not clearly relate to disabled students, though perhaps the respondent's idea here is that the quick formative response afforded by CAA is especially beneficial to some such students. It is often suggested that certain types of adjustment for the benefit of disabled students, for example the striving for clarity of expression for dyslexic students, are of benefit to all students. Pearson and Koppi (2006) say that an important principle of their staff development course on accessibility (disability) issues was "to encourage participants to think not only about making resources accessible but also to consider alternative approaches in the use of online learning to maximise the benefits for all students".

*Q2: the number of questions in the test*

All respondents agreed that there were not too many questions. Half (7) were satisfied with the number of questions, and half (6) thought that there should be more.

*Q3: length of the session at one hour*

No respondents judged the session too long. 8 of the 13 thought it was about right, and 5 said that it could be longer.

*Q4: what you liked most about the test*

The feature commented upon most was the interactive, “hands-on” nature of the test: “people learn best from experience”. This is hardly surprising as few educators today would disagree with the idea of interaction as one of the key factors for successful teaching and learning. Three of the respondents commented on the dyslexia examples, two of them mentioning in particular the “mixed-up letters” and “jumbled words” example. One respondent described the test as at an “appropriate level for staff”, and liked the feedback provided on answering each question and the links to other material. One person liked “the visual interface”, simple as it was. Certain problems with the software surfaced during the test, causing one respondent to say ironically that the bits he or she liked best were “the bits that worked”.

*Q5: what you disliked most about the test*

There was quite a lot of (constructive) criticism expressed here, which is summarised as follows:

- The interface could be better. There was “poor navigation between the questions”, and “design of the pages not very ‘pretty’”.
- 3 respondents felt that there was a “lack of structured introduction to the questions”, “no explanation of what its [sic] for/trying to prove”. It wasn’t clear whether they were asking for an online introduction to the test, or a revision of the verbal introduction, though the former seems more likely in the light of the comment “it would be good to have some warning/info about the test before the first question”.
- On the questions themselves, there were comments on the small number, the fact that they were all multiple-choice rather than of other types, that the questions on hearing impairment were not yet available, that the question on tunnel vision could have been more realistic, that moving text would be good in the visual section, and that “the questions were difficult” – though this last could be considered a positive comment, since the aim of the test was to deliver difficult-to-answer questions.
- The context of the test could be different: “it might be good to use in groups and discuss responses prior to submitting to encourage debate about the issues”.
- The important point was made that the test as it stood did not clearly point the way forward: “not sure how I would use my (newly gained) awareness to build websites and create quizzes”.

*Q6: whether the information given in the session would influence future actions, and if so in what ways*

Two respondents replied “no”, one saying “probably not specifically” and the other (who was clearly the Disability Support expert from Learning Information

Services) saying “I do this already! I don’t think there was enough detail for people to feel they could go away and make changes but at least they are aware”.

The remainder responded with a “yes”, though their responses tended to show a vague and well-intentioned attitude rather than concrete proposals, which clearly reflects on the achievements of the session. Comments raised the following points:

- The respondents would have a changed attitude: “when meeting disabled students... I will be able to empathise with their problems”, “has given me considerable insight into student needs”.
- There was a general, non-specific, intention to “try to use the lessons”, to “try to think about disability issues” etc.
- Particular areas mentioned were colour combinations (twice), font size and type, and question construction.
- Further dissemination of the material to staff was advocated: “I will talk to others developing ALT [assessment, learning and teaching] issues in general”, “I will try to promote this session throughout the INN Faculty and direct lecturers to this information”.
- There was also a reference to helping to progress current research in the disability area.

#### *Q7. Suggestions for improvement, additional features, ideas for new questions*

This section provided much helpful, critical feedback, summarised below.

- The important point was strongly made once more that advice on how to put the principles into practice was needed: “more helpful to have a session which actually shows you how to build a CAA product using the good points you have in the samples”, “make stronger link to people’s everyday work; give examples”, “more examples of good/bad practice which delegates could evaluate”.
- “More” of several features was requested – more questions, more time (especially to follow the links), and more information on other types of disability, for example restricted mobility (requiring the use of a wheelchair) and other “obvious” impairments.
- The use or demonstration of specialist software (assistive technology) such as screen readers was mentioned.
- Some improvements to the set of questions were requested: “the visual and interactive features of the questions could be improved to better illustrate disability issues”, “be good to get the ‘hearing’ section online”.
- One respondent suggested the possibility of “having a student with learning disabilities attend the session to air their views”.

- Another respondent said that the session might be useful for a specialised group, such as a course development team.
- Liaison with a local Further Education College was suggested, as they have “a very good set of teaching tools for teaching people with physical and mental impairments”.
- Making the session available on the Leeds Metropolitan University website was advocated.
- One comment that is fundamental to the way forward for this work was: “it’s too big a subject for 1 hour and can’t cover ‘the basics’ for people who don’t work regularly in this area”.

*Q8: whether you would recommend this session to others*

9 of the 13 respondents said that they would recommend the session to others, two did not respond and two replied that they would not recommend it at the moment.

*Q9: any other feedback*

A few further points were made here, some of them reiterating earlier ones such as the need for “a shortlist of a few simple actions/guidelines/do’s and don’ts that participants can focus on as an ‘outcome’”. Navigational aspects were again raised (“make ‘useful links’ questions open in a new window”), as well as organisational issues relating to the session. These included sharing the contact details of the session presenters and those attending, providing better background information on the presenters and the project, and providing a zip file of the question website available to the participants.

## **Second version and second run**

The feedback provided from the first run of the test provided many ideas for improvement. The timescale before the next test run, which was timed to take place during Leeds Metropolitan’s annual Staff Development Festival in September 2006 (Leeds Metropolitan University, 2006) was relatively short, given staff time commitments in the intervening period. Effort was concentrated on

- improving the interface and especially the navigability
- ensuring that the test was robust and error-free
- adding an introductory screen of explanation about the aims and format of the test
- adding more questions
- improving the set of web links referring to relevant information and advice

The original question interface divided the screen into three sections (see Figure 2). These held the question itself on the left and one or two useful

links at the bottom of the screen. The right hand section contained a short explanation of the impairment illustrated, instructions if relevant on how to view the question in its “impaired” or “unimpaired” form (normally by clicking a button), and sometimes brief advice. The re-designed screen placed the question in the middle of three sections, with instructions on the left and a set of web links on the right (see Figure 4).

Colour Blindness ALT 2006 Question Set		
<b>Instructions</b>  Click here with the mouse if you are having problems and then try again.  <input type="button" value="Normal View"/>  <input type="button" value="Black and White"/>	<b>Question Body</b>  The most common form of colour blindness is between which pair of colours?  <input type="radio"/> RED / GREEN  <input type="radio"/> YELLOW / BLUE  <input type="radio"/> PINK / PURPLE  <input type="button" value="Submit"/>	<b>Useful Links</b> <input type="button" value="Colour Blindness"/> <input type="button" value="Color Check"/> <input type="button" value="RNIB - Web Access Centre"/> <input type="button" value="VisCheck ColorBlind Vision"/> <input type="button" value="Design for Colour Blindness"/>

**Figure 4: sample question test 2**

Several additional questions were created, including two questions on hearing impairment. One of these showed a video of poor practice where lip-reading is being used – the lecturer, who has a beard and moustache, turns away whilst speaking, brushes his hand across his mouth, and generally makes it difficult for the lip-reader. The other question played an audio file giving an example of tinnitus.

A much fuller set of web links was incorporated in the test, in an attempt to address some of the criticisms raised by the participants in the first test. Some of these links can be considered useful in providing constructive advice on how to prepare teaching materials and CAA tests to take account of the needs of disabled students, for example in the areas of dyslexia (CETIS-TechDis Accessibility SIG, 2006) and colour-blindness (Lighthouse International, 2005). However, there is still a considerable need for further thought and development in this area.

The test was this time presented at a workshop session during a Staff Development conference. No fixed PCs were available at the conference venue, so wireless-enabled laptops were used. These generally worked well, though fixed PCs may be preferable as using an unfamiliar laptop can slow down working. The session was an hour long, and was run by the same three presenters as before, and also by Andrea Gorra, a newly appointed Research Officer for CETL ALiC.

## **Evaluation and feedback from second test**

A small amount of feedback was solicited at the end of the session by way of a brief questionnaire, and responses were received for 11 of the 14 participants in the session. The questions were

- did you find this session useful, and if yes, what was useful?
- any suggestions for improvement?
- any other comments?

In general, the participants said that they had found the session valuable in increasing their awareness and understanding of disability. Much of the detailed feedback was on similar lines to that received after the earlier session, though there was a greater emphasis on the need for examples to follow. A typical comment in the suggestions section read “some examples of how these considerations might be applied to a computer-aided assessment in a specific discipline (i.e. before and after)”. There were several references to the need for future developments, for example “work with others in faculty and university to develop more ... session[s]”, and “best followed up with more detailed sessions, e.g. on hearing impairment / dyslexia”. An important point about measuring what the test achieved was raised (though not entirely clearly) in “– are the results of the test then collected [sic] .... – how might students/learners measure overall progress/learning + feedback overall on questions submitted”.

## **Other feedback**

Some additional comments were received from Alistair McNaught of TechDis , (TechDis, 2006a), regarding technical aspects of the test. He was positive in his comments on the overall concept. He also suggested the possibility of adapting and incorporating SimDis the disability-simulation section of the TechDis Web site (TechDis, 2006b), into the test.

## **Perception of improvements required**

Arising from the feedback comments, there appear to be three main areas that should be considered – technical, pedagogical, and presentational.

### *Technical improvements*

Whilst there has been considerable progress between the first and second versions of the test, some minor errors remain, and a revision of the implementation of the test is planned.

### *Pedagogical improvements*

Clearly there is much work to be done here in providing examples of good practice in CAA for disabled students, clear advice on what staff can do, and

possibly (as was suggested by one participant) some “before “ and “after” scenarios. The range of questions could be extended, and the current questions scrutinised for shortcomings.

### *Presentation improvements*

Decisions on whether to continue to present the test as a short face-to-face workshop, whether to offer it as an online resource (which could then be made available outside Leeds Metropolitan University, perhaps to Further Education colleges in its Regional University Network), and whether to expand parts of it into more detailed follow-up sessions remain to be considered. A proposal to run it as a workshop of one and a half hours at the 2007 annual conference of the Higher Education Academy (Higher Education Academy, 2007) has been submitted. There is the larger question of whether the current session length is too short to be able to offer anything sufficiently useful. Pearson and Koppi (2006) discuss a staff development programme whose aim is “to enable staff to develop competence in the design of inclusive and accessible learning resources, to apply their knowledge in the development of their own projects and to encourage other staff to consider accessibility issues in e-learning resources”. Although this has a wider focus than just CAA, it is interesting to note that the two delivery models that they describe for their accessibility course are “the face-to-face one-day workshop; and the flexible online course which is more intensive and may take place over one or several weeks or even a whole semester”. Whether a one-hour course is sufficient to achieve anything useful needs to be reviewed, although one of the initial aims of the project was to provide a session that staff could easily find time to attend.

### **Limitations of evaluation**

So far, evaluation has consisted primarily of feedback from the staff involved. Evaluation of staff development in universities may be undertaken in different ways, for example using the Content/ Input/ Reaction/ Output model (Northumbria University, 2006). A feature of the evaluation that has not yet been undertaken in this study is the “output” phase of such a model, which would consist of attempting to assess the impact of the staff development activity. This could include analysis of the take-up (by measuring the number of staff who choose to interact with the test, and by recording the adoption of it by other institutions such as Further Education colleges), as well as more extensive and structured collection of feedback from users. Possibly a community of users engaging in a dialogue around the test, based on current social networking principles, could be encouraged.

### **Conclusion and future work**

This staff development test to introduce disability issues in relation to technology and CAA has been found to be instructive and useful by its participants. Improvements have been made, but further work remains to be

done in several areas, especially with regard to providing useful models to staff in how they may improve their practice and in the area of assessing the test's impact.



## References

- CETIS-TechDis Accessibility SIG (2006) *Cognitive Disabilities – Web design for Dyslexia*,  
<<http://www.cetis.ac.uk/members/accessibility/links/disabilities/cogdis#web>>  
(12th February 2007)
- Durham University (2006) *Active Learning in Computing*,  
<<http://www.dur.ac.uk/alic/>> (12th February 2007)
- Harrison, G. and Gray, J. (2006) *A Computer-Assisted test for Accessible Computer-assisted Assessment*,  
<[http://www.caaconference.com/pastConferences/2006/proceedings/Harrison\\_G\\_Gray\\_J\\_s1.pdf](http://www.caaconference.com/pastConferences/2006/proceedings/Harrison_G_Gray_J_s1.pdf)> (12th February 2007)
- Higher Education Academy (2007) *Higher Education Academy Annual Conference 2007* <<http://www.heacademy.ac.uk/events/conference.htm>>
- Leeds Metropolitan University (2006) *Staff Development Festival 2006*  
<<http://www.leedsmet.ac.uk/festival/06/index.htm>> (12th February 2007)
- Leeds Metropolitan University (2007) *Active Learning in Computing*,  
<<http://www.leedsmet.ac.uk/inn/alic>> (2nd March 2007)
- Lighthouse International (2005) *Effective Color Contrast, Designing for People with Partial Sight and Color Deficiencies*,  
<[http://www.lighthouse.org/color\\_contrast.htm](http://www.lighthouse.org/color_contrast.htm)> (12th February 2007)
- Northumbria University (2006) *Human Resources, Evaluation of Staff Training and Development: Guidance Notes*  
<[http://northumbria.ac.uk/sd/central/hr/std/td\\_eval/](http://northumbria.ac.uk/sd/central/hr/std/td_eval/)> (12th February 2007)
- Pearson, E.J. and Koppi, A.J. (2006) Supporting staff in developing inclusive online learning. In: Adams, M. and Brown, B. (eds) *Towards Inclusive Learning in Higher Education*. London and New York, Routledge.
- Techdis (2006a) *TechDis* < <http://www.techdis.ac.uk/> > (12th February 2007)
- Techdis (2006b) *Sim-dis: A view into the unknown*  
<<http://www.techdis.ac.uk/resources/sites/2/simdis/index.htm>> (25th February 2007)



# **MEETING RISING STUDENT EXPECTATIONS OF ONLINE ASSIGNMENT SUBMISSION AND ONLINE FEEDBACK**

**Stuart Hepplestone and Richard Mather**



# Meeting Rising Student Expectations of Online Assignment Submission and Online Feedback

Stuart Hepplestone and Richard Mather  
Learning and Teaching Institute  
Sheffield Hallam University  
Sheffield S1 1WB

[S.J.Hepplestone@shu.ac.uk](mailto:S.J.Hepplestone@shu.ac.uk)

[Richard.Mather@shu.ac.uk](mailto:Richard.Mather@shu.ac.uk)

## Abstract

Students at Sheffield Hallam University are increasingly demanding the ability to submit assignments online and to receive feedback and their marks online. A key theme of the University's Learning, Teaching and Assessment strategy (2006-2010) is "to enhance the students' learning experience, making assessment activities, support and feedback a powerful integrated feature of learning". Students will be encouraged to reflect on feedback to "enhance their on-going learner development through timely and effective feedback". This short paper will explore how the University is currently working to meet its students' expectations for online assignment submission and online feedback, through the development of a new Blackboard Building Block that supports the flexible submission of student assignments and the timely delivery of feedback online.

## Introduction

As most student assignments now originate in an electronic format, the ability to submit work and return feedback online offers natural benefits for students (Bridge and Appleyard, 2005) including:

- the saving of paper and printing costs (and postage costs for distance-learning students)
- the flexibility to submit assignments any time, any place
- speeding up the process of returning feedback (as students no longer have to wait until their work is returned to a collection point)

When responding to the Student Expectations survey (in which new and returning students are invited to express their expectations of a supportive e-learning environment) conducted at the University in September 2006, many students commented on the usefulness and flexibility of online submission and electronic feedback, and that they would expect to be able to use it within their modules:

*"My placement is an hour away so being able to submit work online is extremely helpful"*

*"I think all coursework assignments should be sent electronically"*

*"I will be submitting assignments electronically"*

*"Online feedback from tutors is a brilliant idea...I sometimes find that I can forget verbal feedback"*

*"Online feedback is useful"*

When asked in a more recent survey (February 2007) about what enhancements they would like to see made to their existing Blackboard sites to improve their online learning experience, students at the University once again responded in favour of being able to submit assignments and receive marks online:

*"Handing in essays online would be helpful"*

*"Ability to submit work on Blackboard"*

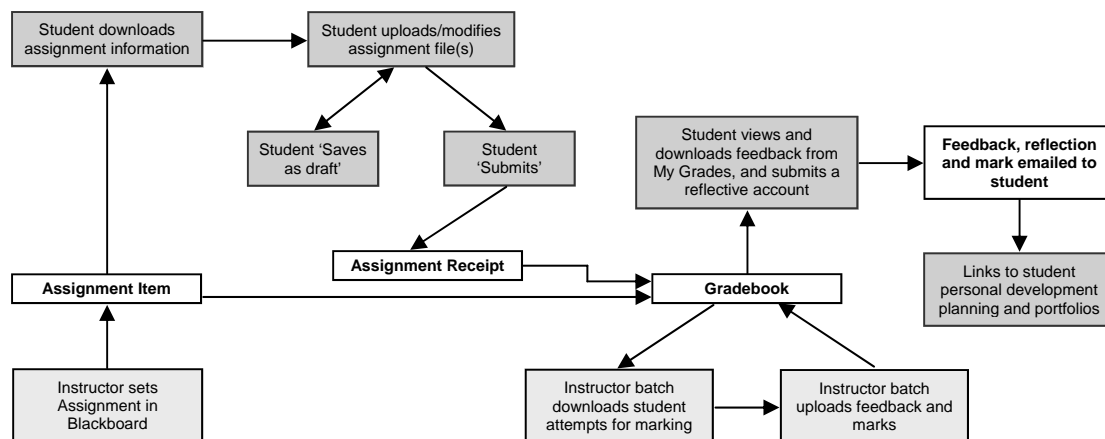
*"More online things like handing in ... coursework"*

*"Being able to receive all grades online, including coursework"*

*"Coursework grades put on them"*

### **The Sheffield Hallam Assignment Handler**

To meet these growing student expectations of online assignment submission and feedback, the University undertook a project to investigate improving the way that student assignments are processed and to enhance the way in which feedback can be provided using Blackboard. Tutors at Sheffield Hallam have been able to receive, track and store student assignments, and return marks and feedback in the Blackboard Gradebook since 2003 using the Assignments functionality. The starting point for this project was to map out the lifecycle of a student assignment from the perspective a tutor setting an assignment, a student completing and submitting their work, the tutor then providing feedback and marks on the student work, and finally the student accessing their feedback and marks for the assignment (Figure 1). As the resulting 'map' clearly indicates where students have responsibilities in the course of completing and submitting assignments, and reflecting and acting upon feedback, tutors are keen to share this representation with their own students.



**Figure 1: Blackboard Assignment Mapping**

The outcome of this activity highlighted two key areas for development.

1. Supporting the timely delivery of feedback online by:

- Enabling tutors to batch upload student marks with file attachments providing detailed feedback for both assignments submitted through Blackboard and for hard-copy assignment submissions. At present tutors uploading feedback for assignments into the Blackboard Gradebook must access each individual submission, and for large student cohorts this repetitive process can take considerable effort and delay the time it takes for students to receive feedback. Tutors wishing to give feedback online for assignments not submitted through Blackboard are currently limited to only providing marks in the Blackboard Gradebook
- Providing feedback on group assignments to each individual in the group, rather than one per group
- Allowing students to access their feedback all in one place and presented within the context of the module alongside learning materials and activities
- Automatic email notification of feedback availability
- Encouraging students to engage with their written feedback and identify key learning points in order to activate the release of their mark (after Black & Wiliam, 1998, who argued that the “effects of feedback were reduced if students had access to the answers before the feedback was conveyed”; and Potts, 1992, who claimed that abolishing grades encourages students to engage with feedback, as they are “obliged to find for themselves value in what they did”, ensuing a richer learning experience). At present any feedback provided to students is linked from the mark that is displayed in the Student Gradebook tool, and students can simply view their mark without accessing the attached feedback. Students will also be prompted to download a copy of their feedback to attach to their submitted reflective account

## 2. Supporting the submission of student assignments online by:

- Enabling tutors to set up the submission quickly and easily (supporting both individual and group assignments) with the assignment brief presented at the point of submission
- Allowing students to submit their assignments any time, any place, and providing a detailed electronic receipt for their work
- Storing all submitted assignments within the Blackboard Gradebook for staff to access whenever they need
- Automatically renaming submitted files with the module code and student number to make them easier to manage
- Reducing time delay and the administrative burden associated with the distribution of student assignments to tutors
- Providing tutors with an at-a-glance check of who has not submitted to identify at-risk students

In conjunction with Blackboard Inc., these enhancements have been developed into a new Building Block. The Sheffield Hallam Assignment Handler has been made available to tutors at the University during Semester 2 2006/7, and even though its use will be evaluated from Semester 1 2007/8, tutor feedback to date has been extremely positive:

*“Uploading grades individually via Gradebook was time-consuming and frustrating but this has been resolved with the new Assignment Handler”*

### **Electronic and automated feedback generation**

The next stage of our work is to further streamline the process of developing and writing feedback electronically, and to investigate the possible range of methods for reviewing and marking student assignments on-screen.

A generic feedback template is being developed which will allow tutors to create feedback documents specific to each Assignment item created in Blackboard, incorporating features that allow student assignments to be assessed quickly and efficiently. This development follows on from the work of a tutor in the Faculty of Development and Society at the University who realised he was re-writing similar comments when marking his students' assignments. By including a matrix of statements in a Microsoft Excel spreadsheet, he found that he could save time in generating and returning individual feedback for each student. Printed copies of this individualised feedback are supplemented with verbal comments when handed back to students. Individual feedback can also be returned to students via email and this has been used by another tutor in the same Faculty since 2005.



A customised version of this spreadsheet used since 2006 by a tutor in the Faculty of Health and Wellbeing at the University, has reduced her marking time for a cohort of 120 students from “six weeks of intensive marking to three weeks of more relaxed marking”, and now finds that it is easier to mark work consistently. The same spreadsheet is now used among the rest of the subject group. However, the feedback files are printed and returned to the students in hard copy.

All of these developments parallel the work of Denton (2001) who developed a technique using a combination of Microsoft Excel 97 and Microsoft Word 97 to generate personalised feedback sheets which include the student's mark, who also reported that such procedures “can make the assessment of work from large groups considerably less onerous”.

The continued development and use of the original feedback spreadsheet template will be accompanied by the creation of an associated tool that will allow the generation of a generic feedback template for an entire student cohort, which has been developed in parallel with the Sheffield Hallam Assignment Handler. This new tool will make use of the student information downloaded along with their assignment attempts from the Blackboard Gradebook, and will allow tutors to create a matrix of assessment criteria (which the students will have received when the assignment was set) in preparation for reviewing, marking and providing feedback on student assignment submissions. As the tool stores the data automatically for each student as it is entered, marking can take place in more than one session. Tutors enter a mark against each assessment criteria, automatically generating a feedback comment and general comments can be added for each student.

When the marking process is complete, the tool automatically creates a spreadsheet file containing marks and feedback against each assessment criteria for each student. Tutors can select whether to keep the total mark for the assignment hidden from the feedback. All files can then be batch uploaded to the Blackboard Gradebook in a single zip file where the relevant feedback file and mark is automatically attached to the relevant student for that assignment. This has a considerable time-saving benefit for tutors with large student cohorts, as they no longer have to access each student's assignment attempt in turn to attach feedback. Students can then access their feedback all in one place and presented within the context of the module alongside learning materials and activities in Blackboard, rather than separately by email. If the total mark for an assignment is hidden from the feedback, students will be encouraged to engage with their written feedback and identify key learning points in order to activate the release of their mark.

In a more recent and separate development, another tutor in the Faculty of Health and Wellbeing at the University has devised an electronic system that uses ‘visual sliders’ for marking and providing students with feedback. As the sliders are linked to ‘development actions’ which suggest how the students could improve on their performance, the need for being specific about marks is removed. Students receive a visual representation of their position on the

marking scale for each assessment criteria alongside the associated development actions. These are currently output using a combination of HTML, Flash and XML data. This development is still in its infancy and opportunities for linking it to the generic feedback template for batch upload to the Blackboard Gradebook are currently being investigated.

There is still some resistance from tutors concerning the on-screen reviewing and marking of student assignments, with a general perception that it requires them to be in a fixed location and reading on-screen for a great deal of time. To address this we will be exploring, piloting and generating user case studies on a range of strategies and techniques for on-screen reading and marking of student assignments in a paperless environment. In addition to investigating the annotation of student work with typed comments and feedback, we will explore the use of digital ink technology to write handwritten comments (which can be digitised by recognition software) directly onto the surface of a tablet PC or digital notepad, retaining the flexibility of traditional marginal comments (Plimmer and Mason, 2006). This investigation will include the loan of a range of hardware and software to tutors, such as lightweight laptops, tablet computers installed with Microsoft OneNote, and recording equipment for the creation of audio feedback. This investigation will reflect the mapping process used for the lifecycle of a student assignment as demonstrated in Figure 1.

## References

Black, P. & William, D. (1998). Assessment and classroom learning, *Assessment in Education*, **5** (1), pp. 7-74

Bridge, P. & Appleyard, R. (2005). System failure: A comparison of electronic and paper-based assignment submission, marking and feedback, *British Journal of Educational Technology*, **36** (4), pp. 669-671

Denton, P. (2001). Generating Coursework Feedback for Large Groups of Students Using MS Excel and MS Word, [online]. *University Chemistry Education*, **5** (1), pp. 1-8. Last accessed 1 May 2007 at: [http://www.rsc.org/pdf/uchemed/papers/2001/p1\\_denton.pdf](http://www.rsc.org/pdf/uchemed/papers/2001/p1_denton.pdf)

Denton, P. (2001). Generating Coursework Feedback for Large Groups of Students Using MS Excel and MS Word. [online] In: Proc. Fifth International Computer Assisted Assessment Conference, Loughborough, 2-3 July 2001. Learning and Teaching Development, Loughborough University. Last accessed 1 May 2007, at: <http://www.caaconference.com/pastConferences/2001/proceedings/j3.pdf>

Plimmer, B. and Mason, P. (2006). A Pen-based Paperless Environment for Annotating and Marking Student Assignments. [online]. In: Proc. Seventh Australian User Interface Conference (AUIC2006), Hobart, Australia, 16-19 January 2006. Australian Computer Society Inc. Last accessed 1 May 2007, at: <http://crpit.com/confpapers/CRPITV50Plimmer.pdf>

Potts, D. (1992). Case study: You can't teach those things to rats. A case for neither grading nor failing students, *Educational and Training Technology International*, **29** (4), pp.296-309

Sheffield Hallam University Learning and Teaching Strategy 2006/10. Internal document



# **A HARDWARE SOLUTION FOR ACCESS TO CAA FOR STUDENTS WITH REDUCED MANUAL DEXTERITY DUE TO ACUTE NEURODISABILITY – A CASE STUDY**

**Colin Heron**



# **A Hardware Solution for Access to CAA for Students with Reduced Manual Dexterity due to Acute Neurodisability – A Case Study**

Colin Heron  
North East Wales Institute of Higher Education  
Plas Coch Campus  
Mold Road  
Wrexham  
LL11 2AW  
c.heron@newi.ac.uk

## **Abstract**

This study was undertaken to assess possible solutions to the issue of acute disability with regard access to Computer Aided Assessment systems. With the vision for the adoption of CAA set out by the Quality and Curriculum Authority approaching reality, it is now increasingly important for inclusion, widening participation and accessibility to be paramount on the agenda of any course team developing assessment strategies. This paper is a review of available solutions for students with extreme mobility differences that can exclude them from mainstream approaches to the control of software interfaces. It describes a bespoke solution utilised in a particular case that has proved to be invaluable in the teaching and assessment of a student with limited manual dexterity due to Cerebral Palsy. The study focuses on editing packages that are used extensively in the assessment of audio media courses, but the solution brokered could be adapted to CAA across many disciplines.

## **Introduction**

The Disability Discrimination Act (DDA) states that, 'It is unlawful for the body responsible for an educational institution to discriminate against a disabled student in the student services it provides, or offers to provide'. Therefore, it is the duty of the educator to design systems and a curriculum that can be accessed by every student regardless of physical difference.

The latest figures available for the Higher Education sector indicate that problems of physical accessibility were experienced by 0.3% of the total cohort of 895,675 students<sup>1</sup>. In the institute where this case study was conducted, the figure is substantially higher with 4% of the student cohort registering some form physical disability that affects mobility. Because of the

---

1 HESA Statistics 05/06 - <http://www.hesa.ac.uk/holisdocs/pubinfo/student/disab0506.htm>

generic descriptors used to compile these statistics, accurate data for specific student needs that match this case are not available at this time. As interest grows in the area due to the DDA, it is envisaged that data will become available to enable greater analysis of the present situation.

Reasonable adjustment can be used to redress many problems that can be faced, but this can prove to be difficult for students with acute conditions if the course is designed to be assessed on software packages that require a high level of manual dexterity.

Audio media courses are examples of routes of study that rely heavily on the production of artefacts on a wide range of software platforms as forms of assessment. This case study will focus on a student's journey through the curriculum of a course that requires a high level of expertise in the use and operation of such packages. The student has restricted mobility and manual dexterity due to the symptoms of Cerebral Palsy. The condition resulted in the student's access to computer interfaces being limited to the fingers of one hand with reduced movement in the arm. This was exasperated by fatigue, which was the result of the effort required to position a mouse or the use of a conventional keyboard. Another problem that became apparent at a very early stage was that the accuracy of mouse position was going to be an issue due to the working screen resolutions required in the software packages.

The main resource for existing research into this area in the UK is TechDis, a project supported by the Joint Information Systems Committee (JISC). TechDis have funded many projects to investigate solutions into a broad range of accessibility issues<sup>2</sup>. In the poster presentation 'Techdis Accessibility Pyramid'<sup>3</sup>, they advise;

“When investigating appropriate solutions for individuals it is important to consider the amount of time and effort required for them to begin using the relevant kit.”

Throughout this study, this advice has been paramount to the formulation of the solution to ensure that the student is not disadvantaged by the learning curve of the technology.

The target software package that will be described throughout this paper is an audio editing solution from Sony<sup>®</sup> entitled SoundForge<sup>®</sup>. The package requires the user to operate a graphic user interface that relies heavily on mouse movement for data selection and access to drop down menus for editing functions. Due to the particular limitations of finger dexterity, the interface was proving to be difficult for the student to master to the degree of proficiency required by the learning outcomes of the assessment criteria.

The majority of tools for assistive control are centred on mouse and keyboard alternatives. A suite of these are available as standard in the Windows Xp<sup>®</sup>

---

2 <http://new.techdis.ac.uk/>

3 Poster presented at ALT-C 2006 (5-7 September, 2006)



operating system whilst other more esoteric solutions are provided by third party companies. This paper will concentrate on the tools that could be of benefit to the subject of the case study, namely, ones that assist with user interaction and with the selection of data on screen.

## **Standard tools**

Although tools for controlling computer interfaces have been developing rapidly alongside those of the major operating systems, accessibility has been mostly focussed on the major software titles due to the combined effect of market forces and demand. However, the very nature of learning differences due to Neurodisability mean that finding suitable tools to match the software package and the particular needs of this specific individual has proven to be difficult. However, in order to highlight the level of research conducted into this particular solution, the author has included a description of some of the more popular access tools that were discounted due to the students specific learning needs.

### *Sticky Keys*

This is one of the standard tools that form the accessibility suite in Windows Xp<sup>®</sup>. It addresses the problem of a user with limited finger dexterity and allows the user to access keyboard shortcuts that are designed for activation with two or more simultaneous key presses. An example of one of these commands would be the standard Windows<sup>®</sup> shortcut for copying data ('Control' + 'C'). For a user with restricted finger dexterity and the use of only one arm, this command could not be achieved. With sticky keys activated, the user can press a modifier key such as 'Control' and have it remain active until another key is pressed. This could prove useful for simple commands such as the example indicated above, but it would prove to be restrictive from both a time and functionality perspective when applied to complex packages.

### *Filter Keys*

FilterKeys is another member of the Microsoft Xp<sup>®</sup> accessibility suite. It is an option that instructs the keyboard to ignore brief or repeated keystrokes. This offers the advantage to a user with reduced dexterity in that it can aid the input time of data by reducing the error rate associated with inaccurate key presses.

### *Track Balls*

The traditional mouse requires that the user performs three tasks at once. This can prove to be disadvantageous to a user with limited dexterity. The user must grasp the mouse, move the mouse and click a mouse button simultaneously in order to complete a function. By design, a trackball allows the user to perform each of these tasks in a non linear fashion, thus allowing for greater control when movement is restricted. The disadvantage of a trackball for the specific needs of this study is the trade off with regards speed when breaking from the linear approach of standard mouse interaction. Because all control interactions will take between twice and three times that of

the standard mouse function, it will impact upon the fatigue experienced by the user during a session.

### **Expanding the horizons of control**

Because of the limitations of the standard tools, further research was conducted into the type of interaction that could be accepted by the target application. Remote control features for the package are well specified as the software is targeted for use in media production studios as a digital recording platform. The standard control language for this form of operation, due to the environment in which it is designed to operate, is the Musical Instrument Digital Interface. MIDI was developed in the early eighties as a serial communication protocol for the control of musical instruments. In order for this to be effective in this application of the software, a control surface would have to be configured that could communicate the necessary MIDI commands to the host computer. Control surfaces of this type are available commercially and can be configured to send pre-programmed strings of MIDI data. After thorough evaluation, this approach was discounted due to the complexity of the programming of the device and also because this technique would be restricted to the one piece of target software.

### **A bespoke solution**

Whilst researching the control messages required to operate the software remotely, a data sheet was compiled of alternative key presses based on the standard multiple key protocols of the operating system.

It was suggested that a custom built keypad that could send macro commands that were strings of data mirroring the ASCII protocol offered by conventional keyboards, could replace the use of the mouse for the majority of functions required by the user. For indicative purposes, table 1 gives an overview of the basic commands for control of the software package complete with the mouse interaction and shortcut equivalent. It was calculated that for complete control of the package at least sixty distinct keys would be required.

<i>Function</i>	<i>Mouse Interaction</i>	<i>Shortcut alternative</i>
Play	Click GUI Button	Shift + F12
Stop	Click GUI Button	Escape
Record	Click GUI Button	ALT + R
Select all data	Click and drag	Ctrl + A
Increase Magnification	Click GUI Button	Shift + Up Arrow
Decrease Magnification	Click GUI Button	Shift + Down Arrow
Save	Click menu item	Ctrl + S

**Table 1: Comparison of functions, mouse interactions and shortcuts**

It was established that programmable keypads were available that can accept user defined macros per key. These are marketed as components for point of sale interfaces in retail environments. For this project, a keypad with 128 keys was acquired that connected to the computer as a standard PS2 device (Figure 1). The keypad is a flat panel that has a facility for sliding a user defined legend beneath a clear vinyl screen. This allowed the user interface to be bespoke for the application.



**Figure 1: Programmable Keyboard**

The keypad is programmed via a host application called ChangeMe. Once the configuration of the buttons is established in the software, the program is uploaded to the non volatile memory in the keypad itself. The keypad then becomes an autonomous piece of equipment that is dedicated to the software in question.

## **Ergonomic considerations**

For the controller to add enough benefit so as to add extra value to the user, the differential in terms of speed must be measurable. For the increase to affect productivity and to aid with the issue of fatigue, the keys must be grouped into clusters that can be reached in a logical manner determined by the intended function of the individual key. Many of the functions of the software package are already grouped in this manner due to the collections within the drop down menus. The menu collections were utilised in the initial design for the interface, with groups such as the file functions located in strips and coloured via the legend to match the functionality of the cluster. This was repeated with the transport functions, to allow for the complete control of the software from the device.

The size of the keypad enabled its position to be adjusted so that the user did not have to reach excessively in order to access any of its functions. This greatly increased the amount of time the user could participate before the onset of fatigue.

## **Further work**

The keypad has scope for use with other software programs due to the fact that it has four distinct layers that can be applied to each key. This gives the possibility of 512 distinct command strings to be saved in its internal memory. As the Windows<sup>®</sup> operating system dictates that most programs share some shortcut commands, the amount of distinct commands required per program is reduced. The amount of functions required for a particular program would depend upon the complexity of its control interface and the expected level of expertise required to be demonstrated by the user.

The legends that are produced to indicate the functions of the buttons are located beneath a vinyl screen and can be easily changed on a program by program basis. This makes it feasible for the keypad to control numerous software packages without reprogramming.

It is envisaged that a web based depository of pre-programmed macros could be assembled for use as a resource for educators across all sectors. This could make access to numerous software packages a reality to users with severe disability.

This idea could be extended further by the adaptation of Access key protocols by the developers of online assessment materials. Although the Access key attribute in HTML has not reached critical mass due to problems with its implementation, the development of assessment materials could benefit from this technology. The problems that have arisen in the standard HTML protocol have been due to the duplication of command strings with existing shortcuts. Because the keypad can generate strings with the ASCII code for up to four simultaneous key presses, this would not be a problem. Development using other scripting languages could utilise a control surface based upon standard short cut commands in a similar way that the Sony<sup>®</sup> SoundForge<sup>®</sup> example

has been used in this case study. It would be advantageous at this point in the development of the keypad as an access platform to establish a set of Access Key commands specifically for computer aided assessment applications.

## **Conclusion**

The keypad has had a profound effect on the student for whom it was designed. It has reaped benefits that go beyond the initial aim of increased accessibility to the software package. Due to the ability to act independently and in a time frame that allows for a more spontaneous approach to the creation of media artefacts, the student has displayed a vast increase in confidence and self esteem.

The keypad and the methodology from this case study have opened possibilities for the adaptation of more software packages. It has been proved that any computer aided assessment which has options for standard ASCII keyboard shortcuts, can be made accessible to students with acute physical learning differences through devices and strategies of this type.



# **DIAGNOSING AND DEVELOPING THE IT SKILLS OF NEW ENTRANTS TO HIGHER EDUCATION**

**Steven Jones**





# Diagnosing and Developing the IT Skills of New Entrants to Higher Education

Steve Jones (Leeds Metropolitan University)  
e-mail [S.R.Jones@leedsmet.ac.uk](mailto:S.R.Jones@leedsmet.ac.uk)

## Abstract

This paper presents an approach to the diagnosis and development of IT skills using Computer Aided Assessment (CAA). It looks at the rationale for the assessment of IT skills and the relevance for higher education in general. It reflects on some of the outcomes of the project and staff and student thoughts on the use of CAA in this context.

## Introduction

This work reports on a pilot project to establish an approach to the diagnosis and development of students IT competencies using CAA. The project was developed from recognition (within the institution) of the increasing diversity in the IT competency of new entrants into higher education and their increasingly varied background and expectations of university education. Furthermore, that an effective mechanism needed to be put in place to ensure that all undergraduates have a core set of basic generic IT skills in the very early stages of their University education. These skills would then create a foundation of these competencies that could be built upon and contextualised within the HE curriculum. A fundamental aspect of the project was to put in place computer based diagnostic testing to enable tutors and students to assess initial competencies, follow a customised learning plan and to track subsequent development.

## Context: the importance of IT skills to new entrants to HE

There is an increasing amount of research in the Higher Education sector regarding students experience during their first year of study. Yorke and Longden (2004) identified four key reasons why students leave programmes of academic study. Two of these reasons can be seen as being within the area of influence of institutions. These are: the students' experiences of their programme and more broadly experiences within the institution of study and, secondly, students' failure to cope with academic demands made by their programme of study. It is for these two reasons that the importance of IT skills to new entrants are worthy of further investigation.

- Students need a set of generic IT skills to satisfactorily undertake their course of study. Student satisfaction and progression are

compromised if they do not have these skills. Equally where students are uncertain of a particular skill set formative assessment and the accompanying feedback is important (Yorke and Longden, 2004)

- IT skills are a prerequisite for e-learning – usage of Virtual learning environments has increased substantially in recent times and staff in institutions are developing increasingly complex learning systems this makes increasing demands on student IT skills.
- Students without a basic set of IT competencies place considerable pressures on support mechanisms within institutions.
- The use of IT applications within a class may be based upon assumptions regarding existing levels of students' IT competency. If these assumptions do not hold, lecturers can unexpectedly find themselves undertaking remedial work with those students in the class who do not have sufficient IT skills.
- There is increasing recognition of the importance of IT skills in benchmarks on key skills, from professional bodies and employers.

Institutions who do not give students the opportunity to improve and update skills may well be compromising these students chance of success.

### **The project approach**

A set of key principles were established that guided the approach the project.

The first principle was based on a study of entrants over a three-year period that showed a large diversity in the level of IT skills and confidence of individuals. It was therefore important to the team that any approach should take this into account.

The second principle stemmed again from a survey of first-year entrants that revealed a range of preferred approaches to acquiring IT skills. Whilst some preferred a classroom environment others preferred to follow approaches such as working from home.

A third principle was that students should only do what they needed to do and avoid repetition. Student feedback confirmed that repetition of the learning of areas that they were familiar with was de-motivating.

### **Computer Assisted Assessment Process Design**

The approach taken was to design an assessment process that fitted in with several aspects of the University environment such as the use of WebCT and the need to have a system that was aligned to the European Computer Driving Licence (ECDL) since the same assessment process could also be used for staff as well as having value added benefits for some students enabling those who wished to do so, to continue on to the ECDL.

Based around this, the assessment was developed in seven separate tests based around the seven module domains of the ECDL. These tests were standalone in that they could be used separately if required. Each test was developed into a series of sub-sections that also echoed the ECDL syllabus for that particular module domain. Within this sub section a question bank was developed from which a selection of questions was delivered at random to the student. The types of questions used were mainly multiple choice, with the question types restricted to those available in WebCT Campus Edition. The pass mark was set at 80% as with the ECDL, but in the later version this was set at 60% for those not taking the ECDL since this was nearer to the baseline level of competency required by tutors teaching at Level One.

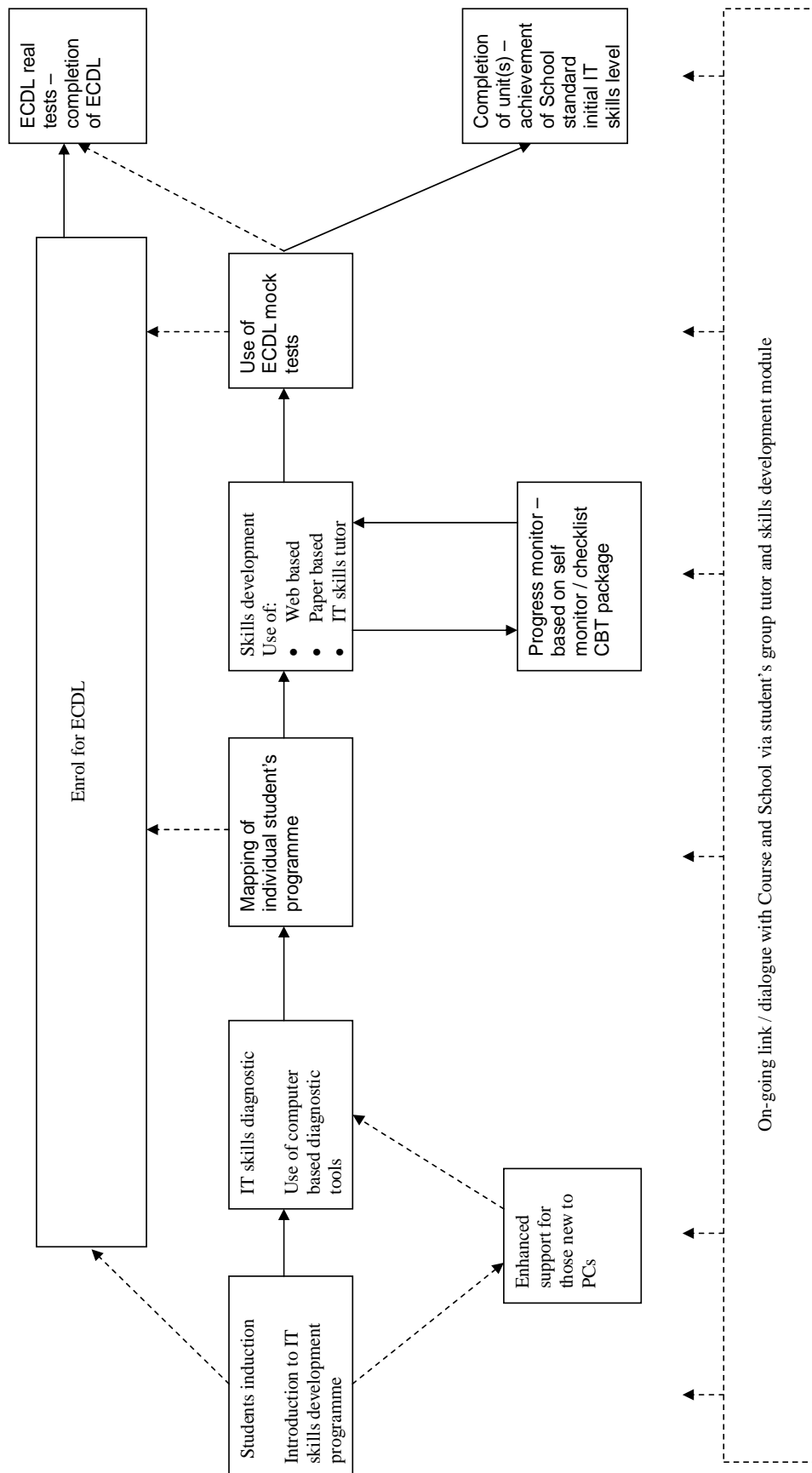
Students were introduced to and sold the idea of achieving a minimum level of competency in IT skills. Initially the students took a pre-course diagnostic questionnaire. The questionnaire established details about how the respondents previously acquired their IT skills and their perceived skill levels and confidence. This was designed to inform approaches to teaching prior to the development of skills and competencies, to provide the student and tutor with guidance on the level of support that they were likely to require and also to inform future developments of this project.

Within the first two weeks of the term students did a series of computer-based diagnostic tests via WebCT. The aims of these tests were to provide an individualised learning plan and to provide feedback to the students on their real rather than perceived abilities using IT. Students were taken through the first two tests (normally Word and PowerPoint) in an appropriate class and shown how the system worked and the resources available.

Feedback was not included with the questions but the students were asked to look at their marks within a subsection and if these did not meet the threshold, the students were asked to work through the relevant sub-sections of the tutorial packages. This was facilitated by the fact that the paper-based and computer-based tutorial packages also complied with the subsections of the ECDL syllabus. On working through the sections of the tutorial packages the student then could re-take the test hopefully to reach the required level or to be in a position to take the ECDL exam.

Marks achieved were monitored and evaluated in terms of overall performance of the cohort taking the tests, something that has led to changes in teaching and initial exercises given to students in subjects that utilise technology. Evaluation of the performance of each individual student was not carried out due to the numbers and time required. Additionally, this was presented to the students as an opportunity and requirement to develop themselves and to reach a baseline level of skills for them to be able to successfully approach modules that use ICT. Students not taking the ECDL have been asked to submit evidence of marks achieved in a skills development portfolio, so avoiding the need to chase up individuals.

Figure 1 shows the process.



**Figure 1 - Key processes**

## **Outcomes**

Feedback on the project came from a combination of student focus groups, general student comments and tutor observations. Five main paradoxes were apparent:

Most new students have some IT experience, but in some this is very limited. Furthermore asking students about their levels of computer experience was a very poor indicator. In particular there was a noticeable difference between genders and their ability to accurately diagnose their level of IT skills. Based on this project team felt that this demonstrated the importance of diagnostic testing of new entrants rather than relying on themselves to position their skills against a set of criteria and take appropriate remedial action.

The students IT key skills and knowledge were patchy. Typically, students who could use programs such as Word, thought they were good using the web to search for information, and said they could use Excel, but few had additional experience. Diagnostic tests revealed that the students could indeed use the Word (but not the more advanced features) but elsewhere had a much weaker skill set. Focus group discussions revealed that many students were used to following a set of instructions to achieve a given goal but were much less adept at thinking of how to apply this to a new situation.

The students recognise the importance of having good IT skills, but did not appreciate the skills they would need for their studies. There was a tendency for students to overrate their skill level (particularly young male students) and to base their ideas on the skills they would need at university on the skills they were taught in school (understandably so). The project team considered that aspects of this may be due to differences in subject areas, something that other authors also noted (Kirkwood and Price, 2005)

Students expressed a range of preferences of how their IT skills training should be delivered. While some preferred a classroom delivery (54%) others preferred more flexible options such as blended modes or entirely computer based delivery.

The students said they were keen to develop their IT skill set but for many this was only if they could gain credit for it and that it was clearly a part of their course. This fits in with Brown, Race and Bull (1999) who state “assessment is the engine which drives a great deal of student learning”. For tutors this was an interesting paradox since students could see the relevance for their future career but were not keen to put the time in on their own.

## **Some reflections on the approach**

The use of computer-based assessment was seen as useful by the project team since it provided a number of benefits.

*From a tutor perspective*

- It provided an independent arbiter of skill levels. This was particularly useful for some students who had high levels of confidence using limited aspects of an application.
- A small number (20 of 394) of students were sufficiently unconfident or inexperienced that they felt it necessary to take up a day long 'IT skills primer' From a tutors perspective in the structuring of early classes it helped to know that these students had (hopefully) some of the basics in place and their confidence boosted.
- Many students had relatively high skill levels using Word but much lower skill levels using programs like Excel. In addition while students had much practice using Word in their pre-HE studies very few had any extensive experience using Excel. One tutor commented "They seem to be great at following instructions to create a spreadsheet, but quite poor when asked to create a spreadsheet themselves"
- Provided an easy way for students to repeatedly assess their progress; this lead to worthwhile improvements e.g. average improvement for using Word was 21% to achieve a mark around the ECDL pass level of 80%
- In some cases, students just took the diagnostic tests to confirm that they had a good skill set. In some modules such as PowerPoint, students scored an average of 62% and found themselves with only limited work to do to reach the 80% level.
- This approach enabled the team to provide repeat assessments and tracking for a large number of students with minimal intervention
- It also enabled fast, relevant and direct feedback to the students, something also noted by Peat and Franklin (2002)
- At the same time this provided useful information for tutors in developing appropriate tutorial support for subjects using IT.
- It achieved greater cost-effectiveness by targeting support and training at those who need it and allowing more effective use of IT labs in the Learning Centres.
- Linking this skills development approach to the ECDL had some problems in that there were some areas of the syllabus that many students fail to see the relevance of. Towards the end of this project links with the ECDL's syllabus was much less emphasised whereas links with the students subject area syllabuses were emphasised
- The option of taking the ECDL and the full seven areas proved initially popular, but enthusiasm for this reduced in the face of work load required. The dropout after one semester was 30.2%. The requirement to study all seven ECDL areas was later reduced to

just Word, Excel, PowerPoint and File Management.

- A few staff not associated with these tests did not see value in testing and further developing IT skills or saw priorities in diagnosing and testing other areas (notably maths and literacy skills)
- This project placed another demand on the congested time around induction and the first weeks of the first term
- The ECDL syllabus is quite large, placing some demands on the students' time for a prolonged time.

#### *Student Perspective*

- Students like being able to practice tests – but disliked the extensive number of questions that were required to diagnose their skills in a given area.
- A number of students commented that they acquired skills by learning from the tests as well as by following the computer based training made available. Charman and Elmes(1998) supported the notion of improvement when CAA is used for frequent formative assessment.
- Students liked the improved choice and flexibility in their learning experience and the catering for different learning styles and preferences inherent in the system. Similarly the freedom to structure their own time and make repeated attempts on assessments was important , something also noted by Grebenik and Rust (2002)
- A number of students commented on the differences between their perception of their ability and the test results. For example some students who were seen as having expertise by their previous college or school found themselves struggling whilst an 'A' level computer studies student who rated in his skills as low found himself in the top 10 out of 300 students.
- Maintaining motivation and commitment in areas of study such as file management was not always an easy task from any students; indeed a good number did not see the relevance of such areas of study.
- The tests were considered by some students as being too long. This was more of a problem when the student didn't reach the threshold after the first couple of attempts; here the repetition was seen as a strong negative factor.
- There was some initial resistance to taking the tests from those students who already saw themselves as being competent.

## **Conclusion**

As more and more students come through to university with greater experience of using computers it might be thought that there is no further need for the development and diagnosis of IT skills. However, the experiences of this project show that students' skills tend to be within narrow areas such as the use of word and often don't extend to commonplace applications such as Excel. The assessment and diagnosis of student skills can't really be left with students either since this project revealed a number of students whose perception of their skills were seen to be at odds with reality.

The experience on this project has been that a well constructed system can give benefits to the student and also or help the tutor at the same time. Certainly the use of such a system needs selling to the students but once in place students can certainly gain. The gains from the computer assisted assessment of IT skills are not just that the student acquires a mark but gets to know how their skills stand in a number of areas around benefits from an individual learning plan to remedy deficiencies in their skills. Furthermore, in allowing students to repeatedly test themselves and practice the tests some student will boost their skill set further.

Tutors benefit in terms of workload from the use of an automated computer assessment and are less likely to have surprises caused by assuming students will have a particular skill.



## References

Brown, S., race, P., and Bull, J (1999) *Computer Assisted Assessment in Higher Education*. London: Kogan Page

Charman, D. and Elmes, A. 1998 *Computer Based Assessment (volume 1): A guide to good practice* SEED (Science Education, Enhancement and Development), University of Plymouth

Grebenik, P. and Rust, C. 2002 IT to the rescue, 18-24 in Schwartz P and Webb G *Assessment: case studies, experience and practice from Higher Education* London: Kogan Page

Kirkwood, A. and Price, L. 2005 *Learners and learning in the twenty-first century: what do we know about students' attitudes towards and experiences of information and communication technologies that will help us design courses?* Studies in Higher Education London: Routledge

Peat,M and Franklin,S. (2002) *Supporting Student Learning: The Use Of Computer Based Formative Assessment Modules*. British Journal of Educational Technology

Yorke, M., and Longden, B. (2004) *Retention and student success in higher education* SRHE and Open University Press



**QUICKTRI**

**A VISUAL SYSTEM FOR RAPID  
CREATION OF E-ASSESSMENTS  
AND E-LEARNING MATERIALS**

**Don Mackenzie & Matthew Stanwell**



# **QuickTrl**

## **A Visual System for the Rapid Creation of e-Assessments and e-Learning Materials**

Don Mackenzie & Matthew Stanwell  
Innovation 4 Learning  
University of Derby  
Kedleston Road  
Derby  
DE22 1GB

### **Abstract**

QuickTrl is a visual graphical interface for the creation of e-assessments delivered by an enhanced version of the powerful TRIADS Professional assessment engine.

The TRIADS Professional System provides a very wide range of item types, scoring options and delivery modes and is designed for use by specialist teams who may wish to develop e-assessments up to the level of a full simulation. The sophistication of the output is however matched by the skills required from the assessment author/developer and the system has proved to be too complex for everyday use by most tutors.

One of the problems associated with the development of so-called 'innovative items', such as many of those developed in TRIADS, has been that it requires an expert developer to update or modify them, thus restricting their more widespread use and making them less suitable for item banking.

Our aim in developing QuickTrl is to provide an intuitive system that can be installed on a tutor's desktop to enable the rapid creation of e-assessments, without losing the richness of interactivity, flexibility of screen design and delivery functionality provided by the TRIADS Professional system and to facilitate the modification and updating of quite sophisticated items without specialist software development skills.

The system will deliver assessments via Internet, Intranet, LAN, CD-ROM or on a standalone machine. Browser delivery requires MS Internet Explorer v6 or v7 and Windows 98/2000/NT/XP and the Authorware Web Player 2004. Compatibility with Vista is currently being tested.

One-button-upload to servers is incorporated into the system. The system will produce assessment zip files for upload into SCORM compatible Learning Management Systems including Moodle v1.6, Kallidus and Oracle iLearning.

SCORM connectivity with the Blackboard gradebook is currently under investigation. QTI V2.\* import of items will be available in version 2.

The presentation will demonstrate the ease of use of QuickTrl.

# **INCORPORATING AVATAR SIGNING INTO ASSESSMENT ITEMS**

**Mhairi McAlpine, John Glauert, Vince Jennings,  
Neil Thomas and Antony Rabin**





# **Incorporating Avatar Signing Into Assessment Items**

Mhairi McAlpine, Scottish Qualifications Authority;

John Glauert, University of East Anglia;

Vince Jennings, University of East Anglia;

Neil Thomas, RNID;

Antony Rabin, RNID;

## **Abstract**

Approximately 50,000 people in the UK, mostly those born deaf, use British Sign Language (BSL) as their first or preferred language (RNID, 2007). The Disability Discrimination Act puts new responsibilities onto examination boards to ensure adequate access for all candidates, however we are aware of ongoing issues and limitations with the current support that is offered to deaf candidates, and have rejected other solutions on the basis of cost and practicality. This project explores the possibility of incorporating encoded sign language into assessment items which can then be signed by an avatar

## **Introduction**

Approximately 50,000 (or approximately 1 in 1000) people in the UK, mostly those born deaf, use British Sign Language (BSL) as their first or preferred language (RNID, 2007). For many deaf people, born with no hearing, British Sign Language (BSL) is their first language. Since natural spoken language is phonetically based, deaf people have less cues to assist in learning to read and levels of literacy are typically several years behind hearing people of the same age (Marschark, 2007). They cannot use word sounds for learning and their poor literacy compared to their hearing peers, puts them at a disadvantage when taking assessment tests in English.

The Disability Discrimination Act came into force for examination bodies in 2006. As such we are now required to ensure that all candidates have equal access to our examinations regardless of any disabilities. We also have a responsibility to ensure that our examinations are sufficiently reliable and that candidates are not advantaged by any support that they may get to access the examination.

## **Difficulties experienced by Deaf candidates**

The SQA allows communications support to be provided for deaf candidates when they take an examination – should they prefer to do the examination through the medium of BSL, however the quality of the interpretation may vary – particularly for examinations where there are technical terms used that the interpreter may not be familiar with. Additionally as BSL is a visual language, interpretation itself may provide the answer to some of the questions that we might want to ask candidates. The Joint Council for Qualifications comments on the role of interpreters: *“Many signs are iconographic, and therefore explain the meaning of the subject-specific word being assessed: for example, the sign for ‘perimeter’ draws the outline of the shape in space, and so indicates that the perimeter is the distance around the outside of the shape”* (JCQ, 20??).

In 2003, the SQA launched a new examination in IT skills called PC Passport. Although the exam could be done on paper, it could also be completed on computer. Centres and teachers of deaf candidates found the computerised assessment very difficult to work with. A lack of highly qualified interpreters means that the centres have had to use non-native signers to translate the items and the randomised nature of the questions and on-demand test construction meant that the signers were not able to have prior access to the questions. Difficulties were also experienced by translators as a number of computer terms are relatively obscure and on occasion very difficult to sign without revealing the answer to the question. Video clips of signed interpretations were suggested, however this was considered expensive and non-sustainable. A new sustainable and standardised solution was called for.

## **An avatar signing solution**

The University of East Anglia has been researching the use of avatars to provide signed support for native BSL signers for a number of years (Elliot et. al, 2004). Funded by the Teaching and Learning Research Programme, the Scottish Qualifications Authority, the University of East Anglia and the RNID are developing standards compliant questions which incorporate the instructions for sign language representation, which can then be passed to an avatar who can sign the question on request.

This has a number of advantages over video production

- Good quality video needs relatively sophisticated recording facilities.
- Ensuring continuity is problematic as materials are updated since the same signer, clothing, lighting, and camera settings must be maintained.
- Stitching sequences together is impractical, requiring a complete re-shoot if minor changes are needed.

- The avatar signing process is done by first recording a video of signs used. The signs are then notated in a form of “writing” called HamNoSys, which records the movements that the signer is making.

### Figure 1: Example of HamNoSys Notation

**Figure 2: Anna, the Avatar used by the project**

This project explores the question of support for automated assessment, using the PC Passport qualification as an exemplar. QTI 2.0 encoded questions are annotated with the SiGML notation. A specially designed delivery vehicle based on the R2Q2 development with incorporated avatar reads these questions and uses the SiGML to sign the question to the candidate.

It is hoped that the research associated with this development will give us an idea about how equitable such provision may be, eliminating the variable quality of current translations; ensuring that students have their questions signed to them using accepted vocabularies and eradicating any possibility of “clues” being given by the interpreter in their interpretations.

## References

- R. Elliott, J. R. W. Glauert and J. R. Kennaway *A Framework for Non Manual Gestures in a Synthetic Signing*, 2nd Cambridge Workshop on Universal Access and Assistive Technology, (CWUAAT}
- RNID (2007) *Deaf Awareness Factsheet* RNID, London
- M. Marschark (2007) *Raising and Educating a Deaf Child*, Oxford University Press, Oxford

# **THE USE OF WIKIS FOR ASSESSING COLLABORATIVE LEARNER ACHIEVEMENT**

**Mhairi McAlpine**



# **The use of Wikis for Assessing Collaborative Learner Achievement**

Mhairi McAlpine, Scottish Qualifications Authority

## **Abstract**

This paper seeks to examine the potential for wikis in assessing learner achievement. There is a widespread recognition that groupwork is a beneficial method of learning, and that assessment is a key driver in determining the learning methods that are employed (e.g. Scouler, 1998; Black and William, 1998). Examining the processes of assessing groupwork and the potential that new technology can bring to this is essential to expanding its use. One new technology, which can be used to assess groupwork, is a wiki – an editable webpage which can track the comments made, plus any discussion which goes on behind the scenes, and log the time/date of contributions.

This paper reminds the reader why groupwork is such an essential part of student learning, how it is crucial that this is appropriately assessed, how the assessment of collaborative student achievement has been attempted in the past and the ways in which emerging technology - with a particular emphasis on wikis - can enable the assessment of a process which has thus far been hidden without high intervention strategies.

The SQA is currently considering giving candidates on Project Based National Courses (PBNs) access to a wiki for recording and presenting their group achievements – recognising this as a medium which encourages groupworking and allows demonstration of skills in a manner which encourages collaboration and conflict resolution.

## **The importance of assessing collaborative student achievement**

People have for years been exploring ways of bringing teamwork to learning through the promotion of groupwork, and examining methods of assessing groupwork in order to promote a desirable backwash effect (Wolf et al., 1991). Anecdotal evidence suggests that while learning through groupwork and evaluation through group assessment are common in Primary schools, this trend tapers in Secondary, and virtually disappears in the latter years of Secondary school. The phenomenon of groupwork transition between primary and secondary schools is now part of a new research study examining how the beneficial effects of groupwork can be sustained within the secondary curriculum (Groupwork Transition Project, 2006). Much of this is due to the difficulties of awarding individual awards on the basis of work which has been done on a collaborative basis, however there is no evidence that the

need for students to learn the essentials of teamwork diminish, in fact if anything these skills grow more important as employment gets closer (QCA, 2004).

This issue has been recognised for many years, and there has been a trend to reward team-working, in particular through the “working with others” Core Skill<sup>1</sup>, which runs throughout the curriculum from Level A<sup>2</sup> of the Scottish 5-14 curricular framework through the National Qualifications curricular framework and on to Higher Education. This primarily embedded skill is assessed through the evaluation of competences where the student is required to demonstrate good team-working attributes which can be observed by others, or to produce material which demonstrates that they understand the principles of good teamwork (SQA, 2003).

Unfortunately, many of these assessments are artificial – they are either assessing the behaviour but not the output, or the output but not the behaviour. However the critical factor of team-working, or group-working is to behave in a manner which boosts the performance of the overall team – something which cannot be assessed unless the output and the behaviour are assessed in tandem. This has always been a very difficult balance to achieve. As Bennett and Cass (1988) point out there is a tendency to evaluate that which is easily measurable, and as much of the evidence of these skills are hidden within the micro-interactions of the participants hence it is difficult to gather objective evidence of achievement.

### **Key Features in the Assessment of Collaborative Achievement**

Three features are important to consider when designing a new assessment: its validity; its reliability and the washback effect that it will have. A valid assessment is one which measures that which it purports to measure (McAlpine, 2001); validity is generally separated out into three elements of construct validity; content validity and predictive validity. A reliable assessment is one in which the same results are gained time after time (McAlpine, 2001), and is generally measured using either parallel tests or repeat tests and noting the correlation between them. The washback effect is designed to “*induc[e] in the education system the changes that foster [the] skills that the test is designed to measure*” (Fredrickson and Collins, 1989), ensuring that the assessment promotes desirable learning methods.

To ensure validity, it is essential to pay heed to its three constituent components: construct, content and predictive validity. To ensure construct validity, it is essential that defined learning outcomes that are being assessed correspond to the underlying traits, knowledge or skills which comprise the domain of learning. Where a skill which is designed to be learned is the ability to work collaboratively effectively, or produce collaborative outputs, ensuring construct validity means ensuring that these are defined as assessable

---

<sup>1</sup> ‘key skill’ in England and Wales

<sup>2</sup> Key Stage 1 in England and Wales



learning outcomes for the task. To ensure content validity, it is crucial that the assessed outcomes correspond to the learning objectives of the task. Where the learning objectives include working in teams, or the participation in group tasks or the production of collaborative work, it is essential that these are directly assessed. To ensure predictive validity for future success in applying the learning it is important that the assessment is situated in such an environment. Where the learning applied is likely to be done in groups, assessment in the context of those groups is advantageous.

Reliability is crucial in ensuring a high quality assessment; however, this has always been the stumbling block of groupwork assessment. There are various methods in use to assess groupwork, however there is a perception that they are overly subjective and that it is difficult to apply consistent criteria to phenomena which are by essence ephemeral. Although communities of practice (Wenger, 1999) can reduce the subjective and promote common understanding to some extent this is a major challenge for assessing collaborative achievement. Some of the challenges involved in groupwork assessment are covered below, such as the freeloader, social loafing and proximal development effects, which pose particular challenges to reliability.

The methods used to assess collaborative achievement need to be evaluated with respect to the types of learning and student behaviours which they promote. There also needs to be evaluation of to what extent a piece of learning or assessment is truly collaborative – respecting and highlighting individual group members abilities and contributions, and to what extent it is purely co-operative, where the learning/assessment is structured to facilitate interaction, but there is no requirement to involve and respect all group members. (Panitz, 1996). Direct assessment of collaborative student achievement will promote such working and learning styles, however there may be unintended consequences – students may feel obliged to be more extroverted than they would naturally be, or nervous of making tentative suggestions in fear of being marked down. As with Schrödinger's cat, the act of observing changes that which is observed, and these changes must be monitored to ensure that the effects are desirable.

### **Issues in using Groupwork in Assessed Learning**

Oxford Brookes University (2002) identifies five advantages of using groupwork,

- that students can develop skills of collaboration and team-working;
- group work can allow students to undertake a greater variety of assignments;
- group work can allow students greater say in what tasks they do;
- students get to know each other, and form working relationships which have benefits beyond the particular group assignment work;
- work done in groups can be more real than work done by a large class,

...while James et al. (2002) note the educational benefits that groupwork brings, but stress that the design of its assessment is crucial to its success.

There are a number of guides of how to design groupwork to elicit the best from participating students (e.g. Davis, 1993; Issacs 2002, Connory, 1988, Watkins, 2005), which have a number of common themes running throughout. The most notable of these are the recommendations that there should be a clear definition of group membership, and the roles and responsibilities within it; that the tasks which the group are being asked to tackle require a level of interdependence from the participants and that the evaluation of achievement is pre-determined and explicit.

There are nonetheless issues associated with the use of groupwork. Perhaps the most commonly identified is the freeloader problem (Kerr and Brun, 1983) – the question of how to assess individuals who make no contribution to the group effort within an assessed group scenario. Issacs, 2002 suggests three strategies to overcome the freeloader problem, however cautions that distinction between situational freeloading (where less able members of the group are unable to contribute) and deliberate freeloading must be made. A closely related issue is that of motivation loss which it is estimated accounts for over half the perceived problems with group work (Morgan, 2002). One possible explanation of this is that those underachieving indulge in social loafing, allowing higher ability or more conscientious group members to shoulder the majority of the work (Kerr, 1983). However, the alternative explanation offered by Dembo and McAuliffe (1987) is that higher ability individuals within a group take charge to reinforce their status, effectively sidelining the rest of the team.

There are also issues around the structuring of the ability range of groups. Vygotski (1978) talked of the “Zone of Proximal Development” as a space in which a learner could perform a task, only if they were given appropriate support at a slightly higher level of ability than they would be able to achieve themselves. However, this may have the effect of advantaging lower ability students, depressing discrimination and consequently reliability. Webb et al (1997) has demonstrated that the assessed performance of lower achievers was raised when in a group with others of higher ability compared with in a homogenous group - although the same phenomena was not found with higher achievers (Dembo and McAuliffe, 1987), raising the question of whether it is possible to assess a student independently of the group in which they find themselves.

### **Traditional methods of groupwork assessment**

There are a variety of ways in which groupwork is currently assessed without technological assistance. Chin and Overton (2005) mention individual reports; group reports; observations and interviews; group presentations; poster presentations; peer assessment of contribution to group and individual exercises (although they caution that this last one goes against the ethos of group work). From these, the most popular direct assessment methods of

collaborative working are however group reports, observations and peer assessment of contribution.

Assessment of the products of groupwork in the form of group reports or presentations is one of the most popular assessment forms. The major problem with this approach is ensuring that people are adequately rewarded for their achievements, in a situation where the process is hidden and (in a successful project) where the roles and authorships of the group participants are obscured.

Issacs (2002) suggests a number of different marking approaches which are implemented in assessing the products of groupwork – he notes different approaches to the distribution of marks including shared group mark; individual mark for an allocated task within the project; student distribution of pool of marks and students allocating individual weightings as popular forms of mark allocation. A shared group-mark is probably the easiest form of marking however it is commonly believed to be unfair due to the freeloader problem (see above) – although can be justified if these are frequent small group tasks so individuals are being assessed a number of times. Individual marks for an allocated task may allow for individual differentiation, but is unlikely to promote group cohesiveness and may be biased according to the task that the student has been allocated. Student distributions and allocations of weightings may be perceived as fair by the students, however require a deal of skill which may not be present and can have undesirable social effects, while peer evaluation can reflect more the social interactions in the group than genuine contribution or achievement.

Groupwork is frequently assessed through observations, either informally or through pre-prepared checklists. Less frequently video is used to capture the group members' behaviour and reflected on later to evaluate their contributions. Although observation is common, it is normally used only for formative and reflective. Observation is sometimes accorded some summative weight based observational checklists, however even on video the reliability of the assessment is low as it is difficult to capture all of the interpersonal interactions that will be happening simultaneously within even a small group.

There are a variety of ways in which the peer assessment of contribution approach can be implemented (see Issacs 2002 for some examples); however the key feature is that some marks are allocated to the group for distribution among the members on their own perceptions of contribution. This has the advantage of facilitating the assessment by those who were actually involved in the development process, and as such have a privileged perspective on which members made what contributions. Caution must be noted though that in the absence of guidance on what is to be rewarded, group members may not always be consistent or valid in their marking. Furthermore, the marking may be swayed by the individual dynamics which operated within the group. However, Race (2001) suggests that the individual dynamics which come into play become one of the major advantages of feedback - suggesting that students giving feedback on an ongoing basis in the

course of the groupwork can compensate for the difficulties that tutors find in giving appropriate and learner centred feedback.

Beyond assessing the products of groupwork however, there is a desire to assess the “softer” skills of teamwork and problem solving. Materials have been developed by Learning and Teaching Scotland (LTS, 2005) to directly enhance these; although by their nature these are situated skills which require a context to function, thus it makes more sense to develop and assess them within that context. Indeed this is exactly the kind of approach which is encouraged within the 5-14 framework and National Qualifications framework (SQA, 1999).

### **Assessment of teamwork**

Process is an integral part of groupwork, but it can be very difficult to assess. The core skill “working with others” occurs throughout the UK curriculum from Level A of the 5-14 curriculum through to Higher Education. Van Der Zanden (2005) has completed a short review of the main methods of assessment used for this Core/Key skill by the awarding bodies of the UK. It would seem that although awarding bodies settle on a consistent model of Internal assessment and quality assurance supported by external moderation, there is some variation in the types of evidence which candidates are required to produce.

**Table 1: Evidence requirements for "working with others" key/core skill**

	<b>SQA</b>	<b>AQA</b>	<b>Edexcel</b>	<b>OCR</b>	<b>WJEC</b>
<b>Candidate Evidence</b> <i>E.g. the candidate writing a statement about how they performed in the group</i>	Yes	Yes	Yes	Yes	Yes
<b>Tutor Evidence</b> <i>E.g. the teacher writing a statement about how the candidate performed in the group</i>	Yes	Yes	Yes	Yes	Yes
<b>Peer Evidence</b> <i>E.g. the other group members writing a statement about how the candidate performed in the group</i>	Possible	Yes	No	Yes	No
<b>Objective Evidence</b> <i>E.g. video/audio presentation/folder of work which is kept and presented as evidence</i>	Yes	Yes	Yes	Yes	No
<b>Interrogative Evidence</b> <i>E.g. Responses to written questions/oral questioning by an assessor</i>	Yes	No	Yes	Yes	Yes

**(Table adapted from Van Der Zanden, 2005)**

As can be seen, the most popular forms of evidence are candidate and teacher created, with peer created evidence much less emphasised. Although centre support material is provided for the working with others core/key skill, it is not clear how much consensus and consistent application there is of the criteria, particularly where the candidate might be less skilled at reflecting on their skills.

### **Introducing Technology into Groupwork assessment**

Collaborative learning is nothing new, and can indeed be traced back to the late 18<sup>th</sup> century, being employed at the University of Glasgow for philosophy teaching (Gaillet, 1994) however what is new is the technology which can be employed to support the process. Access to virtual learning environments, both for staff and students has greatly increased over the last two years (Jenkins et al., 2005). With 81% of institutions using it for collaborative working, it is the third most popular use made of the medium behind only access to course materials (98%) and access to web based resources (90%).

One of the first uses of web-based technology to enable groupwork was its use in online collaborative communication environments. The OTIS Project

(Higginson, 2002) compiled a collection of case studies from people using online collaborative learning environments, a number of whom used them for assessment purposes (McAlpine and Higginson, 2002).

Some of the findings from assessing the online groupwork mirrored that of groupwork practise in off-line environments. Anderson and Simpson, 2005, for example found that

*"A strong ethic of group responsibility was developed - most online tasks were group tasks that required each person to undertake some part of a task that groups had to report on. "*

(Anderson and Simpson, 2002)

This mirrors the work of Issacs, Watkins and others, who suggested that a key requirement of successful groupwork was the interdependent nature of tasks which ensured that group members were forced to work together.

One of the major strengths of an online collaborative environment is its transparency. As everything is recorded centrally, it can all be assessed. It should be noted however, that although interpersonal interactions occur within the medium as well, these are not directly equivalent to face-to-face communication. Graham and Misanchuk (2004) highlight the need for active facilitation of the group in an online collaborative environment as a key determinant of its success. . This aspect was noted by MacDonald (2002) and McKenzie (2002) as a major incentive to participation in an online learning environment. Student evaluation questionnaires revealed that although there was initially resistance to the assessment of online team working (McKenzie, 2005), it did provide an incentive to participation, drawing in the shyer members of the group and guaranteeing the involvement of all students (McDonald, 2002).

It was also found that within an online learning environment, when students felt under pressure from the need to complete assignments, they lessened their participation in the online discussions. The transparency of the system made this immediately apparent so that the balance of the assessment could be adjusted, creating a favourable washback effect (Anderson and Simpson, 2002)

### **How wikis can improve groupwork assessment**

A wiki is a type of website which allows readers/users to add and edit content and is especially suited to collaborative authoring. There are a variety of software systems available in which to create a wiki, as well as a number of popular and well used wikis freely available over the web – the best known being wikipedia– an online editable encyclopedia, part of the WikiMedia foundation.

In essence it is a simplification of the process of creating HTML pages combined with a system that records each individual change that occurs over

time, so that at any time, a page can be reverted to any of its previous states. A wiki system may also provide various tools that allow the user community to easily monitor the constantly changing state of the wiki and discuss the issues that emerge in trying to achieve a consensus about the wiki content

In terms of what the user sees, this varies from wiki to wiki, some more sophisticated wikis make more of the system available to the ordinary user, while others only have this functionality available to a power user.

Wikis are in essence a collection of documents which can be developed collaboratively by a number of authors. As multimedia may be embedded within the pages, the document is not restricted to text, but can also hold images, audio, video and animation. Rather than a group submitting a project on paper with supporting materials, a wiki could be used as both a working and presentation environment, allowing a narrative to be weaved around the embedded artefacts. Additionally, the use of wikis can overcome a variety of other issues which have been identified in the literature on groupwork.

### **Difficulty with tying individual contributions into a “coherent whole”**

Traditional groupwork submissions are frequently a disparate collection of artefacts created by different people and put together as the group submission (James et al., 2002). Most successful examples of groupwork involve tasks which require interdependence of the participants, so that one participant cannot perform unless others, (Issac, 2002). In a wiki this becomes a part of the natural method of working: although it would be possible to divide up the wiki to allow people separate spaces within it, this would have to be an explicit decision, and one against the ethos of the project and the default set up of the software, rather than a natural way of working in the environment. Thus the environment itself encourages the good practice which the educator is trying to develop in the learners.

### **Risk that a subgroup may take over the project**

Dembo and McAuliffe (1987) identified that there was a danger that a subgroup of confident and well-integrated members may take over a project, either deliberately or by default as the other members feel less engaged and/or less able to tackle their monopolisation. This can be overcome by defining an explicit space for dissent to be recorded and acknowledged, for differences of opinion to be aired and resolved. In traditional groupwork, this may consist of set aside time devoted to this purpose, such as circle time, however frequently this is neglected as an irrelevance or a timewaster, particularly where dissent is being expressed. The discussion pages of the wiki can make for this explicit space – where issues surrounding the project can be discussed openly, but without the accusations of taking time away from the project. This will also record any group-member who is feeling unable to contribute and the reasons underlying it.

## **Group members may freeload**

A commonly identified problem within groupwork is that group members may “freeload” – taking credit for the groups’ achievement while they themselves contributed little to it. The collaborative nature of the environment can encourage co-operation within the team members, where everyone can see the joint effort – both in terms of products and also in terms of what each of the members is contributing, making it less easy for someone to “freeload” on the back of other peoples work. If group members do attempt indulge in freeloading, the allocation of marks can reflect this as evidenced through their contribution via the history and discussion pages, so they would not be benefiting from other work. The History page can be used to explore the contribution of each of the group members to the overall, allowing a means of observing the contribution that each of the members has made to the overall product unobtrusively, while the discussion page can shed light on any controversies or differences in view that the group members have had in the development of the project.

## **Resolving a freeloading issue without destroying group cohesion**

One of the difficulties that groups which are suffering from a freeloading problem experience is how to resolve the issue without it destroying the cohesion of the group (James et al., 2002). As mentioned above the collaborative nature of the environment makes it less easy for someone to “freeload” in the first place as the contribution of each of the team members is more visible. If group members do indulge in freeloading, the allocation of marks can reflect this, as evidenced through their contribution via the history and discussion pages. The History page can be used to explore the contribution of each group member to the overall, allowing a means of observing unobtrusively the contribution that each member has made to the overall product; while the discussion page can shed light on any controversies or differences in view that the group members have had in the development of the project.

## **Instant yet Subtle feedback**

Students, particularly at the age at which they attempt school leaving examinations, tend to be self-conscious about both teachers’ and peers’ views of them. This may be a contributory factor to the freeloader/social loafing issues discussed earlier. Hara (1998) talks about the frustration that students experience with online distance learning, in particular, the lack of immediate feedback in the absence of direct interaction with the supervisor. Also, the impersonal nature of the student/tutor relationship tended to make it difficult to follow subtle cues, making students nervous that they were not submitting that which was expected.

Benfield, 2000, has commented also that in terms of threaded discussion lists...



*"...because online comments are written, they tend to be invested with gravity greater than is the case with normal speech. If you 'say' something 'silly' online, it will stay there, for all to see, for everyone to reflect on. And you are reminded of it every time you visit that discussion area... Others may find that the time they get to reflect and compose their comments invests them with a power they don't ordinarily feel in face-to-face communication..."*

(Benfield, 2000)

With a wiki, the feedback provided is relatively quick, but also subtle – if someone feels that you have made a positive contribution, it will be built on, if it was not so helpful, it will eventually be edited out as others improve and develop the document. As the author of a part of the document is not immediately apparent (although available from the history), there is less inhibition about deleting or changing someone's work as it is already integrated into the body, compared to taking out a section that a group member has written which is clearly identifiable as their work.

### **Additional means of Authentication**

A further issue which is frequently raised in the assessment of groupwork is the difficulty in determining who has contributed what. The best methods of authentication of group members' work – labelling the artefacts which they produced – is the least likely to promote an integrated, collaborative product. In a wiki however, this need to explicitly label and claim is sidestepped by the automatic logging inherent in the system. Of course there are still issues with the security of candidate details, and the possibility that candidates may undermine the login system by sharing usernames/passwords. That is always an issue and can only be overcome by emphasising to candidates the importance of logging in correctly.

Furthermore, the social issues which can be faced in a groupwork situation may be lessened through the detachment of artefact and authorship. In a wiki "ownership" labelling is done automatically and unobtrusively and moreover it encourages people to shape and change others' work, yet retain the authorship identity and the roles that people have played in shaping the final artefact. As the authorship data is held separate from the main body of the text, it becomes detached from the participant, hopefully overcoming some of the shyness identified by Benfield. Also, as people are encouraged to shape and edit each others work, the final product becomes more fluid and retains community rather than individual ownership.

### **Conclusions**

SQA is piloting the use of social software in Project Based National Courses, which require the submission of evidence of participation in a group based project. Together with a blog to allow candidates to report and reflect on their learning, groups will be given a wiki as a presentation and working

environment for the evidence generated. The first candidates will be using the system in August 07, and the first assessment of candidates evidence through the medium of a wiki will take place in July 08.

It is hoped that this will provide additional evidence of ownership in order to grade candidates' work more reliably, through the provision of greater assessment information, and validity, by promoting group-working through the medium used to display its product.

Although we are aware that group-working is associated with a number of issues, including social loafing and freeloader syndrome, we believe that using a wiki will allow us not only to identify these phenomena, but also empower the other members of the group to directly challenge others indulging in such behaviour in a positive and non-confrontational way. We also believe that, in providing a discussion space, this medium can present a solution to conflict arising in a project, which may otherwise hamper progress or cause group-members to withdraw from the work.

We are excited by the possibilities that this opens up to encourage collaborative working and explore new assessment paradigms – seeking to expand validity while retaining the reliable of more traditional assessment forms. We will continue to monitor the effects of groupwork assessment both on the subject under consideration and the core skills which underlie it.

## References

- Anderson B and Simpson M (2002) Programme wide online group interaction: developing a social infrastructure in Higginson (ed.) The Online Tutoring Skills E-Book ISBN 0-9540036-5-9. Available at <http://otis.scotcit.ac.uk/casestudy/anderson.doc>
- Bennett N. and Cass A. (1988) The effects of group composition on group interactive processes and pupil understanding. British Educational Research Journal. Vol 15, No 1
- Benfield G. (2000) Teaching on the Web – Exploring the meanings of silence Available at <http://ultibase.rmit.edu.au/Articles/online/benfield1.htm>
- Black P. and Williams D. (1998) Inside the Black box: Raising standards through classroom assessment. London, Kings College, University of London
- Chin P. and Overton T. (2005) Assessing Group Work: Advice and Examples, The Higher Education Academy, Physical Sciences Centre, Primer 6 Version 2, University of Hull Available at <http://www.physsci.ltsn.ac.uk/Publications/Primer/group6.pdf>
- Connory B. A. (1998) Group Work and Collaborative Writing. Teaching at Davis, 14(1), :Teaching Resources Center, University of California
- Davis B. G (1993) Tools for Teaching , Jossey Bass Publishers, San Fransisco Available at <http://teaching.berkeley.edu/bgd/teaching.html>
- Dembo M. and McAuliffe T. (1987). Effects of perceived ability and grade status on social interaction and influence in cooperative groups. Journal of Educational Psychology, 79, 415-423
- Denton, H. (1990) The role of group/team work in design and technology: Some possibilities and problems. Third National Conference. DATER. Loughborough 1990 Available at [http://www.lboro.ac.uk/departments/cd/docs\\_dandt/idater/downloads90/denton90.pdf](http://www.lboro.ac.uk/departments/cd/docs_dandt/idater/downloads90/denton90.pdf)
- Dirkx J. M. and Smith R. O. (2004) Thinking out of a bowl of spaghetti: Learning to learn in online collaborative groups In Roberts T. (ed.) Online Collaborative Learning: Theory and Practice. Idea Group Publishing, Hershey PA.
- Fredrickson J. and Collins A. (1989) A systems approach in Educational Testing, Educational Researcher, Vol.18 No.9

- Gaillet. L. L. (1994). An historical perspective on collaborative learning. *Journal of Advanced Composition*, 14(1), 93-110.  
Available at <http://jac.gsu.edu/jac/14.1/Articles/5.htm>
- Graham C. and Misanchuk M. (2004) Computer Mediated Learning groups: Benefits and Challenges to using groupwork in Online Learning Environments. In Roberts (ed.) *Online Collaborative Learning: Theory and Practice*. Idea Group Publishing, Hershey PA.
- Groupwork Transition Project (2006)  
<http://www.dundee.ac.uk/fedsoc/research/projects/groupworktransition/links/>
- Hara N. (1998) Students' perspectives in a web-based distance education course  
Available at: <http://php.ucs.indiana.edu/%7Enhara/paper/mwera98.htm>
- Higginson C. (ed.) *The Online Tutoring Skills E-Book* ISBN 0-9540036-5-9.  
Available at <http://otis.scotcit.ac.uk/onlinebook/>
- Issacs G. (2002) *Assessing Group Tasks, Teaching and Educational Development Institute, University of Queensland*, ISBN 1864995017  
Available at [http://www.tedi.uq.edu.au/downloads/T&L\\_Assess\\_group\\_tasks.pdf](http://www.tedi.uq.edu.au/downloads/T&L_Assess_group_tasks.pdf)
- James R., McInnis C. and Develin M. (2002) *Assessing Learning in Australian Universities: Ideas strategies and resources for quality in student assessment*, Centre for the Study of Higher Education  
Available at <http://www.cshe.unimelb.edu.au/assessinglearning/03/group.html>
- Jenkins M., Browne T. and Walker R. (2005) *VLE Surveys: A longitudinal perspective between March 2001, March 2003 and March 2005 for higher education in the United Kingdom*, Universities and Colleges Information Systems Association.  
Available at [http://www.ucisa.ac.uk/groups/tlig/vle/vle\\_survey\\_2005.pdf](http://www.ucisa.ac.uk/groups/tlig/vle/vle_survey_2005.pdf)
- Kerr N. L. (1983). Motivation losses in small groups: A social dilemma analysis. *Journal of Personality and Social Psychology*, 45, 819-828.
- Kerr N. L. and Bruun S. (1983). Dispensability of member effort and group motivation losses: Free rider effects. *Journal of Personality and Social Psychology*, 44, 78-94.
- Learning and Teaching Scotland (2005) *NQ – Interactive Learning Materials for Core Skills*  
Available at <http://www.ltscotland.org.uk/nq/coreskills/index.asp>,

- McAlpine M.(2001) The Principles of Assessment, Computer Assisted Assessment Centre, University of Luton  
Available at <http://www.caacentre.ac.uk/dldocs/Bluepaper1.pdf>
- McAlpine M. and Higgison C. (2001) New Assessment Strategies in Higginson (ed.) The Online Tutoring Skills E-Book ISBN 0-9540036-5-9.  
Available at <http://otis.scotcit.ac.uk/onlinebook/otis-t4.htm>
- MacDonald J. (2002) Integrating Online tuition with assessment at the UK Open University in Higginson (ed.) The Online Tutoring Skills E-Book ISBN 0-9540036-5-9.  
Available at <http://otis.scotcit.ac.uk/casestudy/macdonald.doc>
- McKenzie J (2002) Enriching content teaching through long term process based relationships for online learning support. in Higginson (ed.) The Online Tutoring Skills E-Book ISBN 0-9540036-5-9.  
Available at <http://otis.scotcit.ac.uk/casestudy/mckenzie-b.doc>
- Morgan P. (2002) 'Supporting staff to support students: the application of a performance management framework to reduce group working problems', online at <http://www.business.heacademy.ac.uk/resources/reflect/conf/2002/morgan>
- Panitz T. (1996) A definition of Collaborative vs Co-operative Learning. Deliberations on Learning and Teaching in Higher Education  
Available at <http://www.city.londonmet.ac.uk/deliberations/collab.learning/panitz2.html>
- Oxford Brookes University (2002) First Words: Advice for new lecturers, Oxford Brookes University.  
Available at <http://www.brookes.ac.uk/services/ocsd/firstwords/fw26.html>
- QCA (2004) The Key Skills Qualifications Standards and Guidance, Working with others, Improving own learning and performance and problem solving. QCA London.
- Race P. (2001) A briefing on self, peer and group assessment, Available at [http://www.heacademy.ac.uk/resources.asp?process=full\\_record&section=generic&id=9](http://www.heacademy.ac.uk/resources.asp?process=full_record&section=generic&id=9)
- Scouller K. M. (1998) The influence of assessment method on students' learning approaches: multiple choice question examination versus assignment essay Higher Education Vol 35 No 4 pp 453-472

- SQA (1999) Automatic Certification of Core Skills in National Qualifications  
SQA, Glasgow ISBN: 1 85969 382 2  
Available at <http://www.sqa.org.uk/higher-still/coreskills/CORE.PDF>
- SQA (2003) Core Skills Framework: An Introduction. SQA, Glasgow  
Available at  
[http://www.sqa.org.uk/files\\_ccc/CoreSkillsCombined\\_0103.pdf](http://www.sqa.org.uk/files_ccc/CoreSkillsCombined_0103.pdf)
- Van Der Zanden. L. (2005) A comparison of methods of assessment of Core skill "Working With Others" in UK examination boards, SQA Internal Paper
- Vygotsky, L. (1978). Mind in Society. Cambridge, MA: Harvard University Press.
- Watkins R. (2005) Groupwork and Assessment in HE Academy Economics Network (2005)The Handbook for Economics Lecturers  
Available at <http://www.economicsnetwork.ac.uk/handbook/groupwork/>
- Wenger, E. Communities of Practise, Learning, Meaning and Identity. Cambridge University Press, Cambridge
- Webb N. M., Nemer K., Chizhik A., and Sugrue B. (1998). Equity issues in collaborative group assessment: Group composition and performance. American Educational Research Journal, 35(4), 607-651.
- Wolf, D., Bixby, J., Glenn, J. G., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), Review of research in education (No. 17). Washington, D.C.: American Educational Research

**FULLER, RICHER FEEDBACK,  
MORE EASILY DELIVERED, USING  
TABLET PCS**

**Paul McLaughlin, Wesley Kerr, Karen Howie**





# **Fuller, Richer Feedback, More Easily Delivered, using Tablet PCs**

Wesley Kerr, Karen Howie and Paul McLaughlin  
(paul.mclaughlin@ed.ac.uk).  
Institute of Molecular and Structural Biology,  
Michael Swann Building, Mayfield Rd,  
Edinburgh, EH9 3JR. 0131-6507060

## **Abstract**

We have developed a method to use tablet PCs to enable markers more efficiently to give written feedback on students' work. Comments may either be made in handwriting, or may be typed, or may be presented in type following handwriting recognition. Additionally, any comments so made can be stored and reused, allowing for editing. Importantly, feedback can be made richer by including forward links for students to follow up on common mistakes that they have made so that their engagement with feedback is more constructive. Such feedback would otherwise be very tedious to provide if marking on paper was used exclusively.

We have run this system successfully for two years to mark essays in a large class of 450+ students, using twenty markers. This volume of work was efficiently handled and involved no paper. Checking of marks and assuring consistent standards was much more easily done than with paper.

We consulted students and markers. Students take the system in their stride. They are well able to provide essays, with diagrams and figures. Markers fell into a number of groups. We have learned that there are a variety of marking styles and developed the software to accommodate these. The only software required is Microsoft Word and Excel.

## **The problem addressed**

Good quality feedback is the most single powerful influence on student achievement in higher education (Hattie, 1987). But a number of surveys with students shows that satisfaction with feedback on assessment is the least of all areas considered. (Hounsell et al, 2005, Krause et al, 2005, Surridge, 2006, Hounsell et al, 2007). Several reasons contribute. Too long a gap between submission and feedback is detrimental and a source of dissatisfaction (Gibbs & Simpson, 2004) Crook et al (2005) have evidence from focus groups that students sometimes simply cannot read a marker's handwriting. They also found that students considered tick sheets and/or boxes in which the marker makes comments to be too formulaic.

As Crook et al point out, many of these problems stem from a rise in student numbers that are not matched by a proportionate rise in staff, such that marking becomes a burden not a teaching opportunity. Marking and returning work for large classes indeed takes much time and resources, both for academic and administrative staff. Traditionally this is done on paper, which has the drawback that handing it back to students causes problems. Hounsell (1987) shows that many students don't pick it up. It sometimes goes missing (perhaps maliciously). If it is collected marked work often goes into a drawer, or is otherwise misplaced, such that the student can't find the work when preparing for a subsequent essay. Submission of word-processed work onto a Virtual Learning environment (VLE) might seem to solve many of these problems, but it creates others. More discursive work, such as the traditional essay, is frustrating to read on-line as most screens are not large enough to display an A4 page at sufficient resolution. This entails much tiring scrolling. Even if the marker has a large enough screen it is rarely portable, and so doesn't fit in with the way most markers work with paper copies. Marking on line also means that feedback must be typed. This becomes very tedious and especially frustrating in the sort of exercise where the marker often has to make much the same comment on many students' essays, or a make a comment that is only slightly edited from student to student

### **A proposed solution**

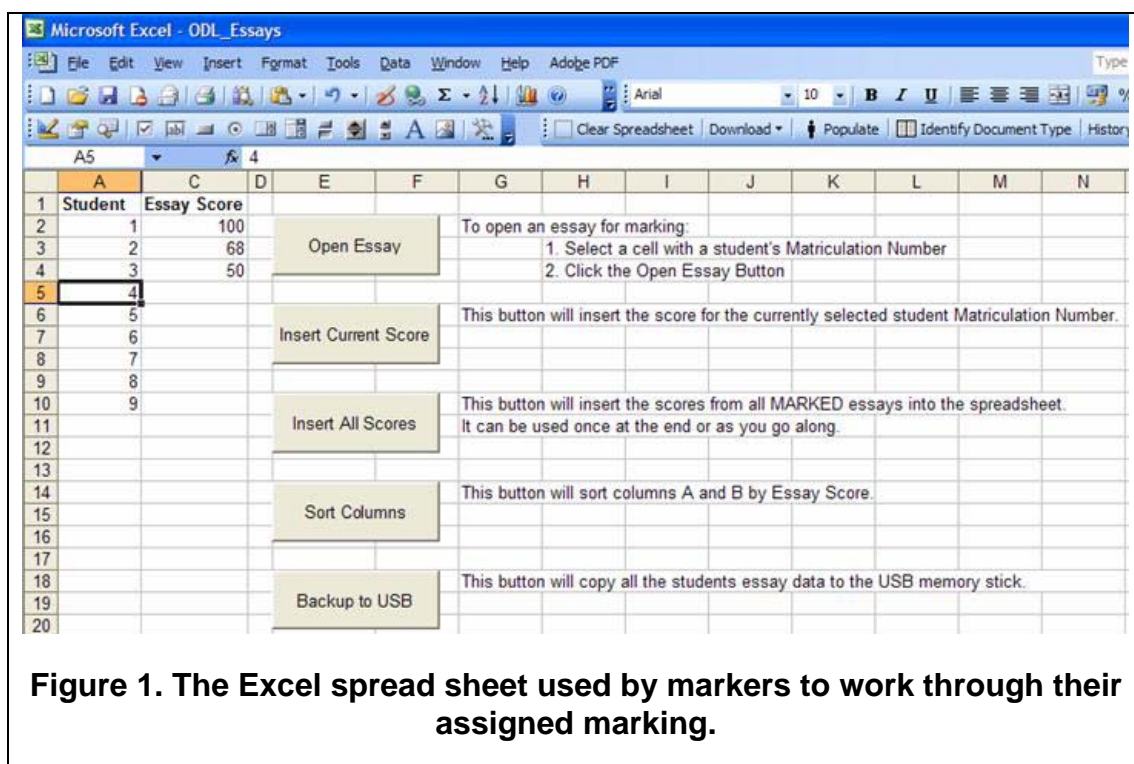
This paper shows an attempted solution to some of these problems using two features of tablet PCs. These machines look like ordinary laptop computers, except that the screen can be swivelled to lie flat such that the keyboard is hidden underneath. Then the screen can display in portrait mode, as opposed to the usual landscape view, such that the screen is similar in size to a sheet of A4 paper. Indeed, a page of a Microsoft Word document can be displayed a page at a time at sufficient resolution to be easily read. and figures are similarly as readable as on paper. The second unique feature used is that the tablet is supplied with a stylus that can be used to write on the screen. The stylus effectively annotates the displayed document in "virtual" ink, again at sufficient resolution that it appears to be similar to writing on paper. Importantly, under Windows XP Tablet operating system there is handwriting recognition such that the writing input by the stylus may be converted into text. Using Microsoft Word Macros, we developed these features into a system to mark submitted work.

### **Implementation – first iteration**

The software has been developed and used to mark essays in a large first year biology class in the University of Edinburgh. The class has roughly 480 students, each of whom submits an essay on a topic associated with evolution. The task is designed to promote students to find material to support their arguments, to help them to appreciate the differences between what is expected at school and at university, and to challenge the misconceptions that many still have about evolution (a pastiche would be

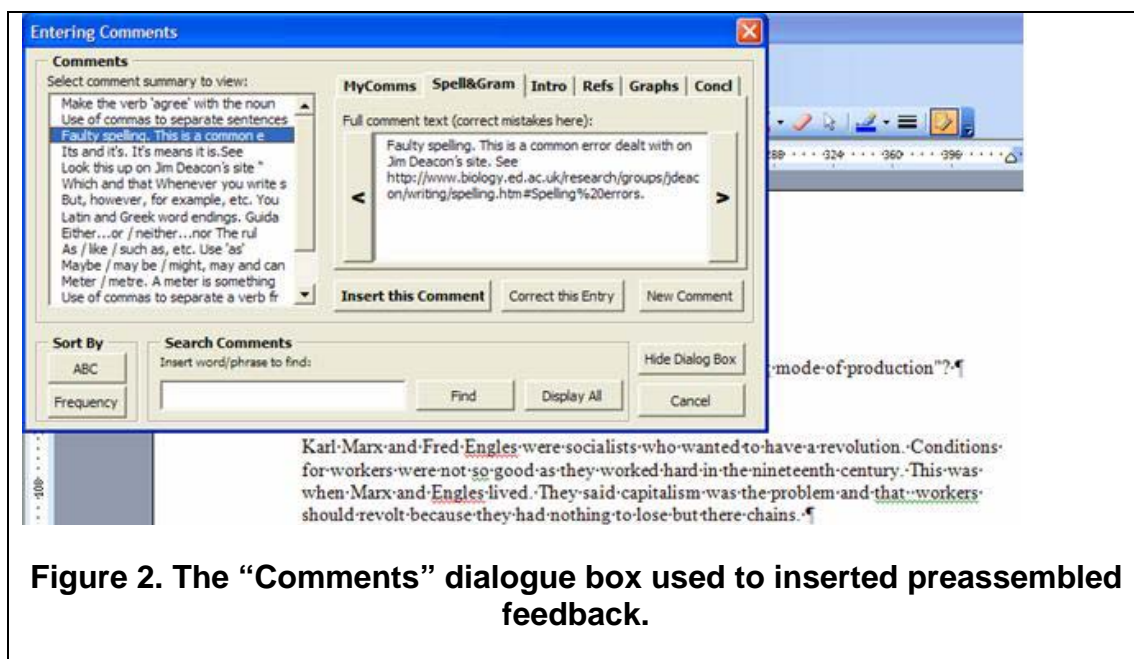
“Giraffes grew long necks to be able to eat leaves on tall trees”, but often the argument appears in essays in a more subtle form).

After completing their essays, students load them, containing their associated diagrams and figures, as Microsoft Word files onto a VLE (WebCT). These are bundled into zip files and downloaded onto Tablet PC machines, which are distributed to each marker. A “Shortcut” icon on the desktop takes the marker to an Excel file, which control the work flow. A macro button populates the file with a list of students.



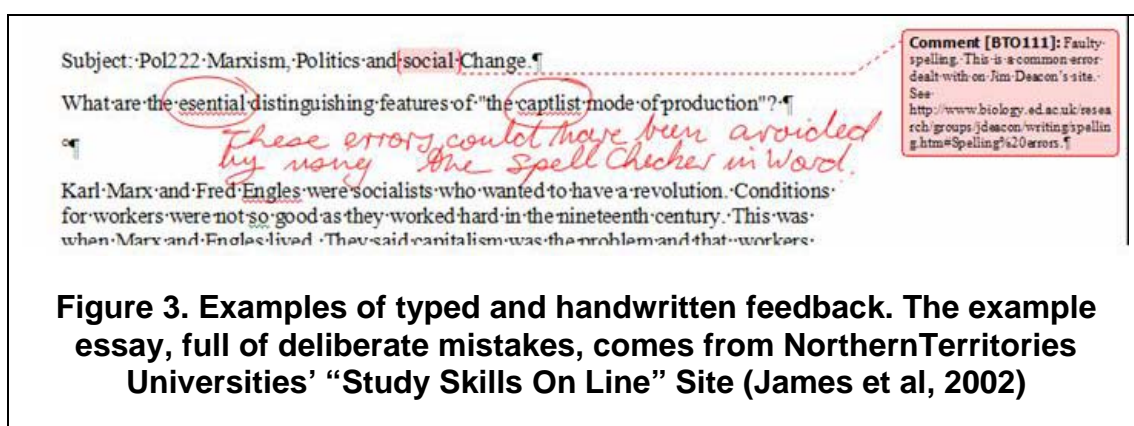
**Figure 1. The Excel spread sheet used by markers to work through their assigned marking.**

Each student may be selected (Figure 1), and on pressing another macro button, the student’s file is opened by calling an instance of Microsoft Word. For the purposes of marking, a particular template has been developed with a number of new toolbars and macros to facilitate marking (Figure 2). The most significant is a “Enter Comments” toolbar, which allows the insertion of comments, their storage and/or reuse.



**Figure 2. The “Comments” dialogue box used to inserted preassembled feedback.**

Previously stored comments may be sorted by frequency of previous use, or may be searched by keyword. They may be edited again to provide a more appropriate comment for a particular student. These comments are inserted as in “balloons” in the right hand margin, as they use the same “Comments” tool provided in Word.



**Figure 3. Examples of typed and handwritten feedback. The example essay, full of deliberate mistakes, comes from Northern Territories Universities’ “Study Skills On Line” Site (James et al, 2002)**

The pre-stored comments are an opportunity to provide students with links to remedial action. Like many HE institutions, the School of Biological Sciences in the University of Edinburgh has a website on generic skills, such as the elements of writing essays, a site on statistics and a site on spelling and grammar. Some of the preassembled comments have links to these inserted. The idea here was to both publicise these sites for students, and also hope

that a student who needed correction on any one point might be drawn to other content via the link.

When the student's work is opened a mark sheet is automatically appended with the marker's name written in and with a space for a mark or grade. When the marker finishes marking, the file is closed. The marker then returns to the Excel spreadsheet and may push a macro button that causes each marked file to be visited, any mark to be read, and then the mark to be inserted into the Excel file. Thus the marker learns how far through the list they have progressed. By this method they are less unlikely to miss an essay out, or misplace a mark than if they had to transcribe marks themselves.

When all markers have returned their machines, their original Excel files are ignored. The directories containing the marked essays are bundled into a new directory structure. Then a similar Excel file reads all students' marked files, the marks and markers' name into itself. Again, marks are read from the files that will be returned to students. This is an important point because it deals with a situation whereby a marker might update the mark on the essay, but forget to update it on the Excel file. Ultimately the marker's own sheet is only of relevance to the marker to track where they are in the list of students to be marked: the student's marked essay is the "golden copy" always.

A master Excel spreadsheet controls all subsequent administration. It is used to look at markers' averages and is used to prepare new bundles of marking to be reassessed by more experienced markers for those markers who have egregious averages. It is also used to make new bundles to be reassessed for those students who are borderline fails, or so that the work of students who were not known to be special needs at the time of marking can be revisited. The spreadsheet is also populated with submission dates so that lateness penalties can be flagged. Those students who attract penalties for plagiarism are also flagged.

Finally, when all work that should be reassessed has been returned, the marked work is moved to a secure website that is protected by the university's authentication system. A dynamic link is released to the students that parses the directory name from their User Identification on WebCT and allows the student to access their own marked work and no-one else's. Such systems are not essential. It would be relatively simple to modify the Excel code to send the marked work by e-mail.

### **Implementation – second iteration**

After the first year roughly 1300 unique comments that had been created by markers were available. The subject of the essay changes every year, so it was desirable to have only generic comments (468) to be used in subsequent years. It was decided to divide these into folders, to reduce the length of each list. The folder names were: General Comments; Spelling and Grammar; Introduction, References; Graphs, which included comments about diagrams, figures and graphs, and Conclusions. After removal of almost identical

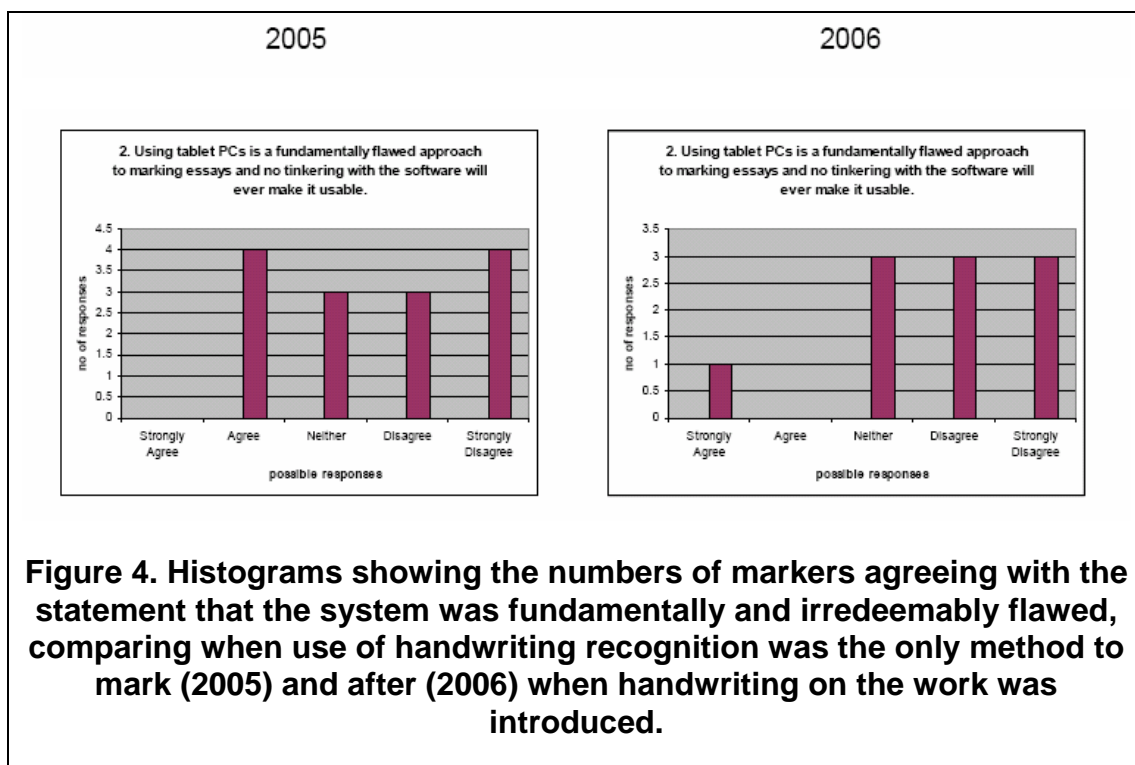
comments, but still allowing different ways of saying the same thing, the total number of comments was roughly 160.

For the next year handwriting onto the essay was introduced in addition to typed on handwriting recognition. In the first year, all comments were made by typing or by using handwriting recognition. From feedback from markers it was clear that some markers found this frustrating. Therefore in the second year we introduced markers to using the stylus directly onto the submitted work. Additionally, in the first year some markers found scrolling with the stylus to be frustrating as the mapping from the stylus to the vertical scroll bar at the edge of the screen was not accurate enough – it was also frustrating for left-handed people who found stretching across their own field of view to be annoying. Thus we decided to buy a mouse with a scroll-wheel for each machine, and this seemed to eliminate these complaints.

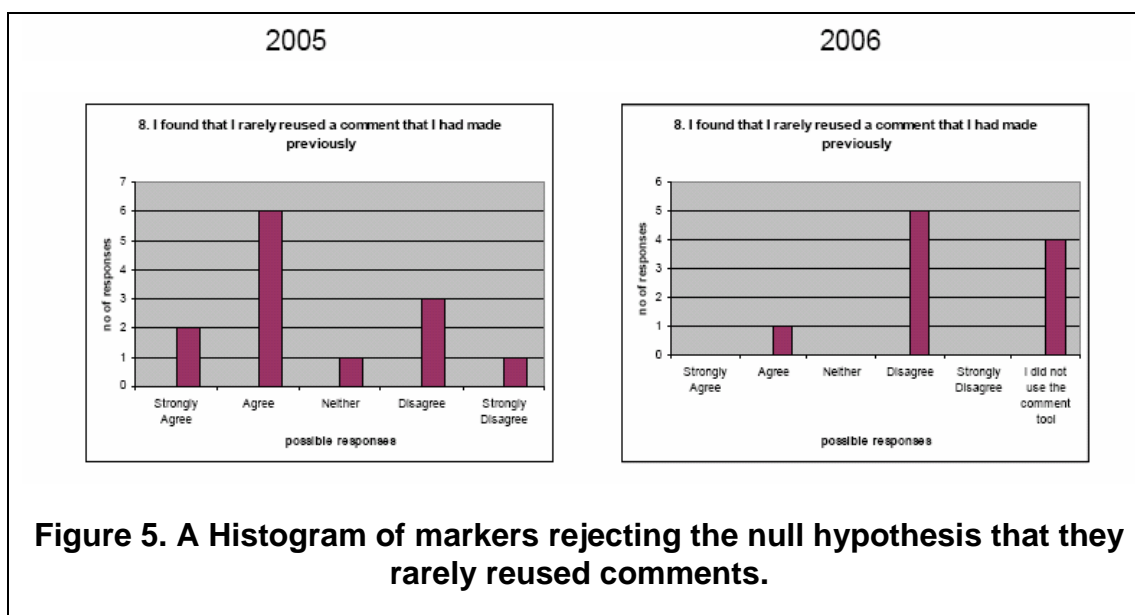
## **Evaluation**

In the first year we paid particular attention to the markers' experience. It was they on whom the greatest burden of dealing with new and unfamiliar software fell, while for students little new demands were made. The most significant difference between this first iteration and the second is that we informed the markers that they could handwrite on the essays, whereas in the first year we led them to believe that handwriting recognition was the only way they could make comments. This was a deliberate deceit because we wanted to capture all comments in machine-readable form so that we could build up a database of comments to form a new list of generic comments for the next year. A second reason was that using handwriting recognition is initially slower than handwriting. We wanted to see if markers would progress in their skills at handwriting recognition and we felt that if an easier option was given, many would not persevere.

It became clear that a significant number found handwriting recognition very frustrating and that marking roughly twenty essays each was not long enough to make sufficient progress. In the second year, 4/10 markers who replied made comments exclusively in their own handwriting, while the other 6 used handwriting recognition or a mixture. Where specific markers chose to identify themselves, there was no clear correlation between either computing confidence or age with use of handwriting exclusively. Some unconfident users, who in the first year complained bitterly about handwriting recognition, used it exclusively in the second without protest. On the other hand, some younger tutors preferred handwriting exclusively.

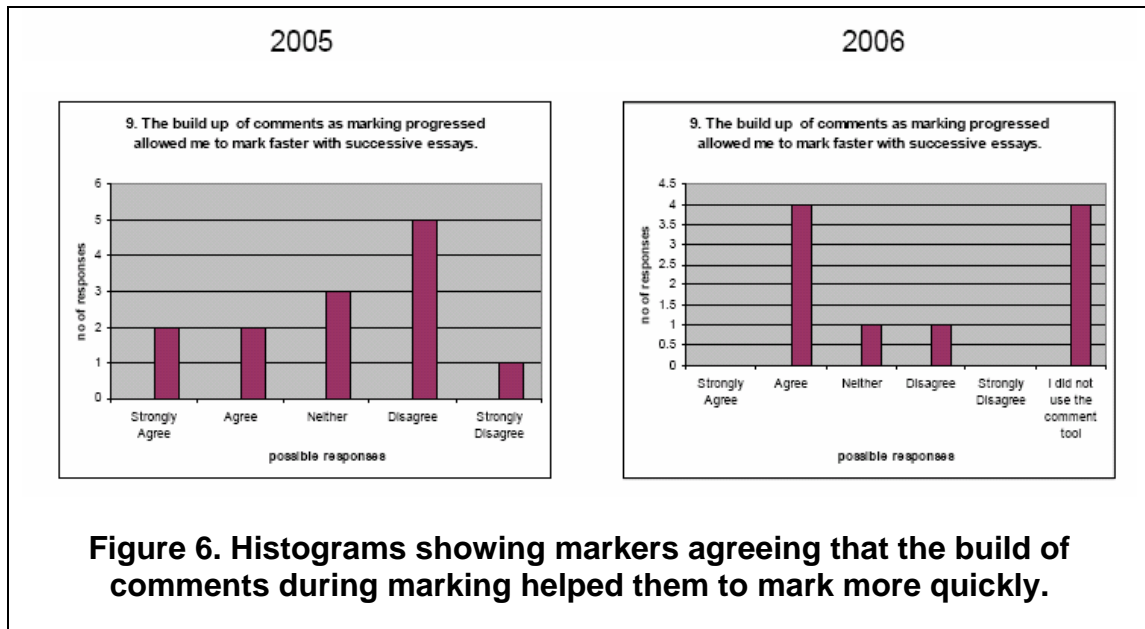


After handwriting was given as an option, there was a significant reduction in those that agreed with the null hypothesis (Figure 4), namely that the exercise was a “fundamentally flawed approach to marking essays and that no amount of tinkering with the software will ever make it useable”



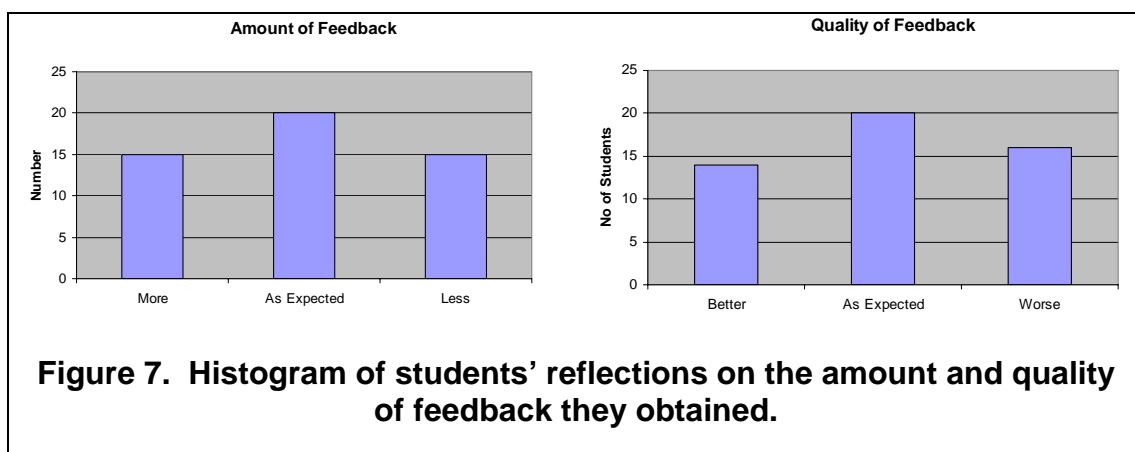
There was also an improvement in those that reused comments (Figure 5). This might have been because in the second year there was a richer bank of

generic comments to use, derived from real comments made by the markers themselves in the previous year. In the first year preloaded comments were sparser and were invented abstractly rather than based on marking real essays. It might equally well have been that they did not use the comments tool bar.



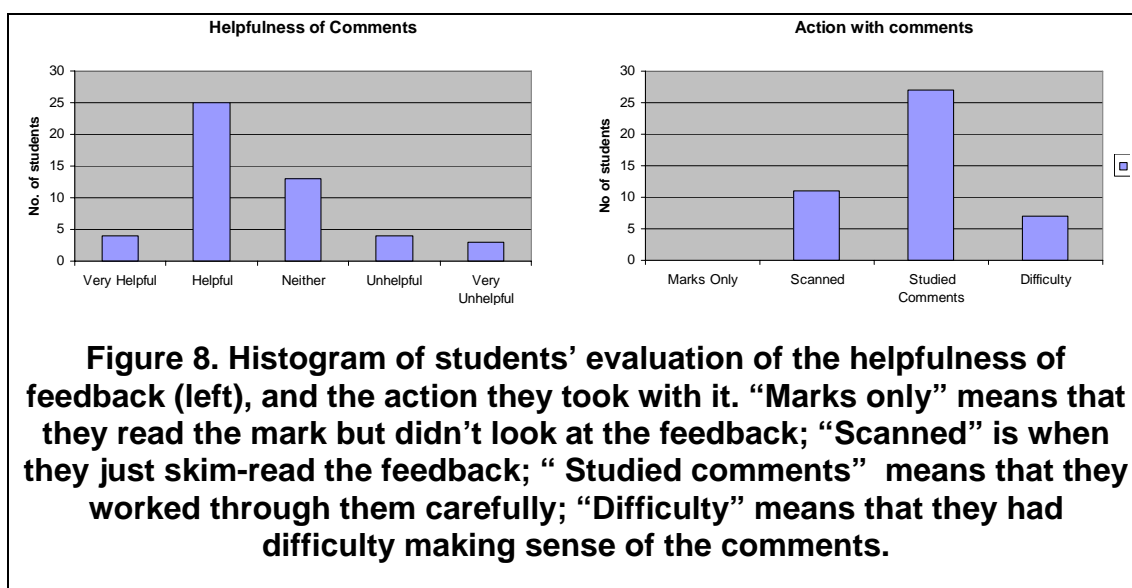
It was also clear that there was an improved perception that the build up of comments during marking was improving the speed of responding as marking progressed (Figure 6).

In the second year, we conducted a survey of student reactions to the essay feedback. We had no baseline to compare improvement against. Not for the first time, a technological development leads to wider reflection on what was normal practice before the innovation. More students had their expectations of the amount and quality of feedback met or exceeded than were disappointed.





However, it should be stated again that these were first year students and this essay was the first many had done at University. Thus these data do not disentangle the effect of the technology from orthogonal factors, such as their expectations from school or disappointment in their grades.



The students' perceptions of the helpfulness of the comments were more positive than negative (Figure 8). The diligence with which the students studied the comments seems gratifyingly high, although we only have their own word for this. Some had difficulty with understanding comments, either because they found the content too sparse or the meaning too elliptical, or because they simply could not read handwritten comments. There did not seem to be much gross difference between the distributions for students whose marked work had typed only comments, handwritten only comments and those that had a mixture. However when the replies were broken down in this way, there were not enough replies to be sure of seeing subtle but significant differences.

We also asked for more discursive feedback from students.

Is there any comment you would like to make on the feedback of your ODL essay?

Individual comments showed that those that were unhappy were usually dissatisfied for a reason unconnected with the technology. For example, there was a perception that it was realistic to attain a mark of 100% and that a marker's role was merely to take marks away, rather than award them. By the same token, we cannot ascribe any positive comments to being solely due to the technology.

- *"Feed back was more detailed than expected"*

- *“Very good feedback, very helpful as a different style is required for a scientific essay than I was used to, so good instructions on this have been given”*
- *“Excellent feed back, very helpful would like [another electronically marked exercise] also to have been marked in this way”.*

There was a clear theme that handwritten comments were often hard to decipher.

- *“It would be good to make all the comments typed, or at least make the markers write in capitals, as I could not decypher the handwriting of my marker.”*
- *“..couldnt read some of the comments made. handwriting was too difficult to read.”*
- *.” I assume the marker used a piece of equipment that allowed handwritting to be shown in a word document. i have re-read the feedback and still cannot make out some of the comments. The feedback given maybe very constructive but i have no way of knowing.”*

We have no data to say that these comments might also have been made on a paper version. The quality of handwriting on the electronic version does not seem to be lower resolution. In the next implementation, however, we will make the default for handwriting a “Biro” rather than a broader-nibbed “felt-tipped” pen.

There were surprisingly few comments on technical difficulties. A worrying case was:

- *“The feedback should also be available to download on macs”*

The College of Science and Engineering at the University of Edinburgh is predominately Microsoft based and Mac computers are in a minority, such that we would have had difficulty accessing a Mac computer to test. An obvious solution will be to save the marked essay as a PDF file.

In summary, it was clear that markers were now much more enthusiastic and positive about using PC tablets now that the software catered for a variety of marking styles. Just as students have a variety of learning styles, markers also have preferred ways of working. Any successful marking engine must cater for these because it is crucial that all take part. The software is still new to many of them and it will be interesting to see how the use evolves and if handwritten comments decline in favour of typewritten ones, particularly as experience grows.

## Conclusions and Perspectives

The system eases administration of large quantities of feedback and should narrow the time gap between submission and receiving feedback. More, but not all, feedback is typed. As the majority of feedback becomes typed issues with legibility should reduce. Also, typed feedback can contain hyperlinks to remedial material on grammar, spelling, dealing with data and the like. A disadvantage of the system, however, is that it relies on tablet PC machines. It would be much more widely applicable if it worked on any Windows machine. Certainly, larger LCD screens that allow an A4 page to be read in one screenful are becoming cheaper, so the need for a portrait screen is lessened. Cheap graphics tablets are available but we have found none that captures handwriting at sufficient resolution. If this problem will resolve itself in the future, the present system could be run with any machine that has Microsoft Word, provided that handwriting recognition software could also be used.

In implementing the system it has been crucial to bring staff along. To this end it is important to develop software that is as flexible and as non-prescriptive as possible. Few academics like to be told that they can no longer do something that they are used to doing. We relied on the goodwill of our colleagues to take the system up. Thus we made it clear that no preloaded comment need be used and if it were used that it should be editable by the marker. Similarly the marker was free to make his/her own remarks and to store them for future use. For the same reasons we eventually “allowed” handwritten comments as well as handwriting-recognition and typing. Interestingly some markers who were vehemently against handwriting recognition in the first year, when it was the only mode of entry, used it in the second in the knowledge that they could have handwritten comments if they had wanted to.

As feedback from students showed, we are not the first to propose a computing solution to find that there are deeper pedagogic reasons for the problem we hope to solve but find that at best we can only mildly alleviate. But this method of marking assists reflection on practice because, by its nature, it accrues large amounts of data that would have been tedious to collect if we had used paper only. Thus we have in machine readable form data to sift for examples of good practice. This might have more weight with markers in the knowledge that it comes from their peers. Future areas to look at are whether novice markers are helped by having a database of remarks that more experienced markers have used, and whether this makes marking more consistent. It would also be interesting to research if more experienced markers feel more of a social pressure to give fuller feedback now that the remarks they make are stored and are seen by their peers, not only by the student being marked. Clearly it is impossible that technology in marking has a neutral effect, but not all changes are necessarily worse.

## References

Crook C, Gross H and Dymott, T (2006) Assessment relationships in higher education: the tension of process and practice *British Educational Research Journal* **32**, 95-114

Gibbs G & Simpson C (2004) Conditions Under Which Assessment Supports Students' Learning. *Learning and Teaching in Higher Education* **1**, 1-31

Hattie, J.A. (1987) Identifying the salient facets of a model of student learning: a synthesis of meta-analyses, *International Journal of Educational Research*, **11**, 187-212

Hounsell, D. (1987) Essay writing and the quality of feedback, in J.T.E. Richardson, M.W. Eysenck & D. Warren-Piper (eds) *Student Learning: research in education and cognitive psychology*, Milton Keynes: Open University Press and Society for Research into Higher Education

Hounsell D et al (2005) *Enhancing Teaching-Learning Environments in Undergraduate Courses: Final Report to the Economic and Social Research Council on TLRP project L139251099*, Universities of Edinburgh, Durham and Coventry: Enhancing Teaching-Learning Environments in Undergraduate Courses Project [www.tla.ed.ac.uk/etl/docs/ETLfinalreport.pdf](http://www.tla.ed.ac.uk/etl/docs/ETLfinalreport.pdf) (accessed 11/05/2007)

Hounsell D et al (2007) Integrative Assessment. Managing assessment practices and procedures. Guide no 4  
<http://www.enhancementthemes.ac.uk/documents/IntegrativeAssessment/IAMmanaging.pdf> [accessed 11/05/2007]

James, R McInnis C and Devlin M (2002) Sample Essay "LearnLine", Charles Darwin University, Northern Territories, Australia.  
[www.learnline.cdu.edu.au/studyskills/as/Sample\\_essay.pdf](http://www.learnline.cdu.edu.au/studyskills/as/Sample_essay.pdf) [Accessed 10/05/2007]

Krause K et al (2005) *The First Year Experience in Australian Universities: Findings from a Decade of National Studies*, Melbourne: University of Melbourne, Centre for the Study of Higher Education  
[http://www.dest.gov.au/sectors/higher\\_education/publications\\_resources/profiles/first\\_year\\_experience.htm](http://www.dest.gov.au/sectors/higher_education/publications_resources/profiles/first_year_experience.htm) (accessed 11/05/2007)

Surridge P (2006) *The National Student Survey 2005: Findings*, Bristol: Higher Education Funding Council for England  
[www.hefce.ac.uk/pubs/rereports/2006/rd22\\_06/](http://www.hefce.ac.uk/pubs/rereports/2006/rd22_06/) (accessed 11/05/2007)

# **IMPLICATIONS OF PATTERNS OF USE OF FREELY- AVAILABLE ONLINE FORMATIVE TESTS FOR ONLINE SUMMATIVE TASKS**

**Jan Meyer, Mel Ziman, Sue Fyfe, Georgina Fyfe,  
Kayty Plastow, Kathy Sanders and Julie Hill**



# Implications of Patterns of use of Freely-Available Online Formative Tests for Online Summative Tasks

Dr Jan Meyer<sup>1</sup>, Dr Mel Ziman<sup>2</sup>, Prof Sue Fyfe<sup>3</sup>, Ms Georgina Fyfe<sup>3</sup>,  
Ms Kayty Plastow<sup>1</sup>, Dr Kathy Sanders<sup>1</sup>, Ms Julie Hill<sup>1</sup>

<sup>1</sup> The University of Western Australia  
Crawley, Western Australia, 6009

<sup>2</sup> Edith Cowan University

<sup>3</sup> Curtin University of Technology

[jmeyer@anhb.uwa.edu.au](mailto:jmeyer@anhb.uwa.edu.au)

[m.ziman@ecu.edu.au](mailto:m.ziman@ecu.edu.au)

[S.Fyfe@curtin.edu.au](mailto:S.Fyfe@curtin.edu.au)

[G.M.Fyfe@curtin.edu.au](mailto:G.M.Fyfe@curtin.edu.au)

[kjplastow@anhb.uwa.edu.au](mailto:kjplastow@anhb.uwa.edu.au)

[ksanders@anhb.uwa.edu.au](mailto:ksanders@anhb.uwa.edu.au)

[jhill@anhb.uwa.edu.au](mailto:jhill@anhb.uwa.edu.au)

## Abstract

The use of online assessment tasks in a summative context can create tensions between the institution's need for security to ensure the validity of individual evaluations and the student's need for flexibility of access. This is especially the case in recent years, with the upsurge of students engaged in paid employment while enrolled in full-time study. The lowest rate of engagement of students in paid employment at the three institutions in which our study was based was 65%, the highest 75%. One quarter of all students at this institution spent more than 20 hours a week in paid employment. Ninety seven percent of students in paid work were enrolled on a full-time basis.

This study determined from automatically recorded times of logon, individual question submission and whole test submission the patterns of use of online feedback-enriched MCQ tests by 656 students across the three institutions in Perth, Western Australia. The conditions under which the tests were available to students varied from a strictly secured, summative task available for a limited time on campus within hours governed by the accessibility of automatically locked-down computer rooms and the availability of staff for live or video invigilation to a freely accessible formative learning exercise.

Mismatches between preferred and available times severe enough to exclude some external students from assessment were identified. Evidence was found that for younger (16-18 year old) students especially, meaningful engagement with test-structured tasks lasts no more than 10 minutes, one third of the designed time of our current summative online tests. The one third, approximately, of enrolled students who did not use the online test facility had significantly poorer academic outcomes. The advantage granted by test use increased substantially with repetition.

The question of how to ensure the security and validity of online testing while increasing real flexibility of access remains unresolved for us. We accept the social responsibility of finding a solution.

## **Introduction**

Barbara Stewart (2004) argues persuasively the potential benefits of online learning and assessment for meeting the needs of underserved populations, such as those with physical handicaps, variant personal cognitive and psychological orientations, who are subject to geographic and cultural separation, and operating under gender and occupational constraints. Curtis and Shani (2002) reported an increase in the proportion of students in paid employment from a single department in a British University from 43% to 55% between 2000 and 2001. Levels of participation in paid employment by first year students at Australian universities have increased from 51.3% in 1999 to 54.9% in 2004 (Krause et al., 2005), at the same time as the proportion of students with language backgrounds other than English has risen.

Stewart goes further than pointing out the benefits of online course material, arguing the social obligation to provide flexible access to learner-centred and assessment-centred learning environments. The use of online assessment tasks in a summative context can create tensions between the institution's need for security to ensure the validity of individual evaluations and students' needs for flexibility of access, however. This study examines patterns of use by students of freely-available formative and constrained summative online tests in an attempt to gain some insight into the magnitude of the mismatch between students' needs and preferences and the current manner of presentation of summative online testing in one area of scholarship at three universities in Perth, Western Australia.

## **Materials and Methods**

The data analysed in this study arose from the development and evaluation phases of a trial of automated explanatory feedback comments for single topic online MCQ tests in first year Human Biology units at three West Australian Universities. Approximately 2,000 students in all enrol in these units each year. Demographic information was gathered from 1099 of these students in a survey of perceptions of the adequacy, use of and need for various types of feedback administered at the outset of the courses. The patterns of use were determined from automatically recorded times of logon, individual question



submission and whole test submission for 656 students across the three institutions. Additional demographic information together with self-reports of test scores was gathered from an online survey of 315 students at the end of semester. Final grades and the contribution of different sections of the courses to those grades were taken from unit exam databases.

The project from which this study arose involved the fitting of feedback comments to online test systems already in use at each institution. Thus, while it was not possible to implement a balanced experimental design, the different situations in which the tests were used and the different characteristics of the platforms through which each was presented did present opportunities for gathering information in a form best likened to that arising from a hybrid cross-sectional/ longitudinal study. For example, it was possible to link self-reported test marks with actual test marks at one institution, expected scores with actual scores at another and the time taken over each question with sectional course marks at the third.

At the first institution the online multiple-choice style tests with feedback were only presented summatively, contributing 24% to the final mark for the course (6% per topic test). Students had 40 minutes to complete the 30 item test at a pre-booked time between 9am and 5pm in a secure, invigilated central computing facility. At the second university the test was presented only as a formative learning task, freely available for 24 hours a day for one month. This test comprised 50 unvarying questions. At the third university the test was presented as a freely available formative exercise for 24 hours a day for one week before items from the same test bank were presented in a summative task available from 9am to 5pm under video surveillance in a departmental computing laboratory. The 30 questions in this test were selected randomly from sets of between 5 and 15 alternatives. The hours of availability of the summative tasks were in both cases governed by the accessibility of automatically locked-down computer rooms and the availability of staff for live or video invigilation.

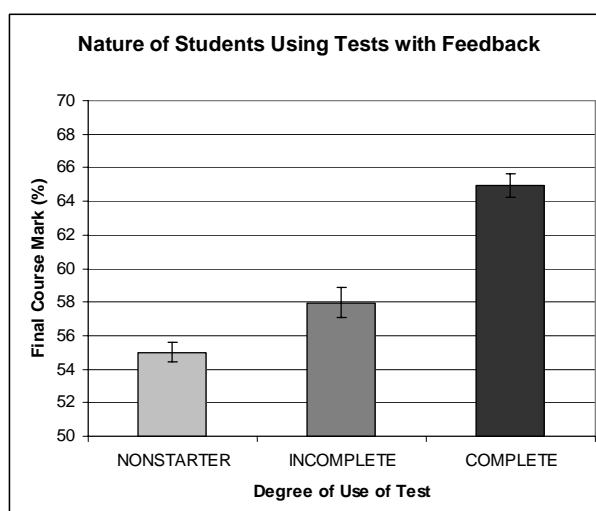
The differences in final course grades of students taking complete, incomplete and no online tests were evaluated using a 1 way ANOVA, with institutions as blocks. At most 6% the final grades used to determine the type of student making use of the online tests was determined by scores obtained in the test (and that for 23% of students). The advantages given by involvement with online testing were assessed by comparing percentage multiple choice question scores in the topic areas dealt with by the feedback-enriched tests in final course examinations in 1-way ANOVAS, with scores in other topic areas as covariates. All analyses were carried out using GenStat ninth edition (2006) and graphs prepared through GenStat and Excel.

## **Results**

The lowest rate of engagement of students in paid employment at the three institutions was 65%, the highest 75% at the institution offering the test as a formative learning exercise only. One quarter of all students at this institution

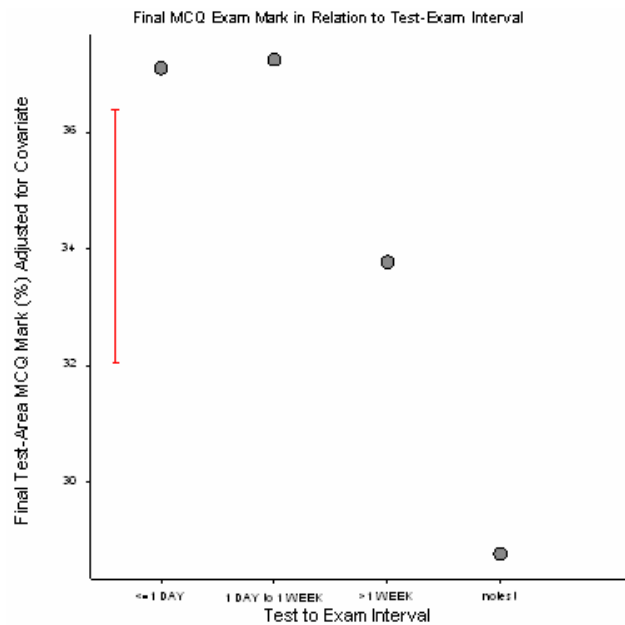
spent more than 20 hours a week (maximum 70hrs) in paid employment, largely in paramedical areas. Ninety seven percent of students in paid work were enrolled on a full-time basis. Twenty two percent of students spoke one of 42 languages other than English at home, nearly half exclusively so.

Rates of use of the online feedback-enriched tests were significantly lower in the institutions where they were not obligatory (63.4% of enrolment compared with 71.9%,  $\chi^2=8.44$ , 1df,  $p=.006$ ) and in the institution where available for a week (60.3%) compared with a month (67.5%,  $\chi^2=4.50$ , 1df,  $p=.034$ ). The students who made use of the online tests were the higher achievers, whether or not those taking the test as an obligatory summative task were taken into account (Figure 1)( 1-way ANOVA, with Institutions as blocking term  $F = 60.39$ , 1 & 1167df,  $p<.001$ , each level differing significantly from the others at  $p <.05$  by LSD).



**Figure 1. The final course grades achieved by students according to the level of use of the online feedback-enriched tests. Marks from the online tests in question contributed at most 6% to the final course grades of fewer than one quarter of the students. (mean  $\pm$  SE)**

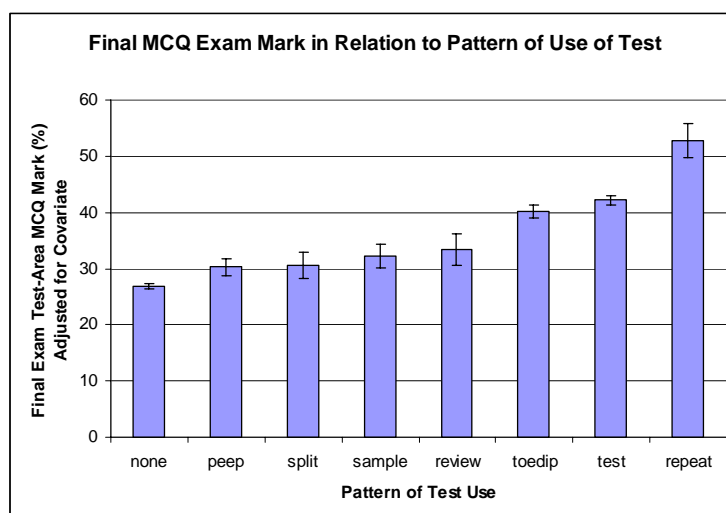
In the institution in which the test was available as a formative review before summative testing use peaked on the days immediately preceding the summative tests. The students who used the test as a review in the week before the summative test gained more advantage from its use than those who had used it earlier, or not at all (Figure 2) (ANOVA  $F= 4.82$ , 1 & 362 df,  $p=.003$ ).



**Figure 2. The advantage gained from use of feedback-enriched tests in relation to the interval between test use and exam at institution using tests in both formative and summative tasks.(MCQ means in test topic area adjusted for scores on non-test topics as covariate, bar = SE of difference)**

The average time taken by students to read the MCQ stem and the five answer options, decide upon a response, review their grade and read the one feedback comment they received was just under 45 seconds. Only 20% of students spent more than a minute on each question, 56% took between 30 seconds and a minute and 24% less than 30 seconds.

It was only possible to analyse the pattern of interaction of students with the test over time in the institution in which it was offered both formatively and summatively. Only one third of the approximately 60% (281) of the class who logged on to the test at least once, completed it once, straightforwardly from beginning to end. Nearly 40% of those logging on never completed a test, 4% did a little more than one complete test and 9% repeated the whole test (one student 27 times). Students who repeated the test showed significantly greater advantage in the final MCQ exam in the topic covered by the online test in relation to other topics than other students (Figure 3) (ANOVA  $F= 4.89, 7\&459df, p <.001$ ).



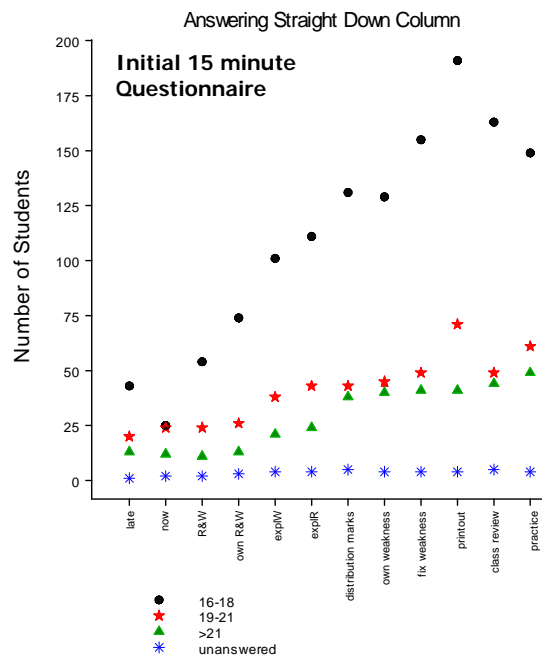
**Figure 3. The advantage gained from use of feedback-enriched tests in relation to the degree of engagement with the tests at institution using tests in both formative and summative tasks. None = did not take test; peep & sample = incomplete tests; test, review & toedip = between 1 & 2 tests completed; repeat = 2-27 tests completed. (MCQ means  $\pm$  SEs in test topic area adjusted for scores on non-test topics as covariate)**

The remaining 14% remained logged on for the whole of the interval between the posting of the test and the final examinations, completing a few more questions every few hours or days until they had completed the whole test. There were many comments made in the follow-up evaluation survey to the effect that a 30 question test was too long, for example

“Also this was a very long test for a computer test after a while at staring at the computer you start to lose concentration.”

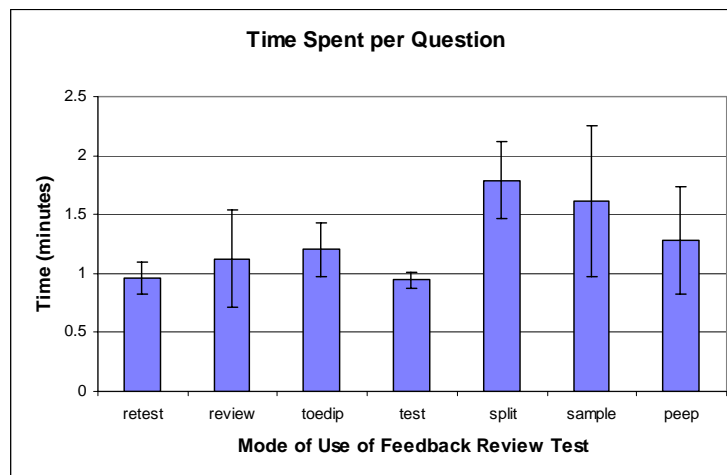
“Big long sentences get too frustrating to read through and understand so often just picked a letter to save time”

Consideration of the rate of fall-off in discriminatory answering of items on the initial questionnaire survey also indicated that younger (16-18 year old) students in particular experienced difficulty in maintaining concentration on a task for more than 7-10 minutes (Figure 4).



**Figure 4. The fall-off meaningful responding across a 15 minute questionnaire, defined by the selection of the same response from the top to the bottom of the column of questions. The first column on the questionnaire is indicated on the far left and the last on the far right. The age groups indicated are 16-18 year olds (circles), 19-21 year olds (stars) and over 21 year olds (triangles).**

The only students who did spend more than the intended minute on each test question were those who split the 30 question test across several sessions (Figure 5).



**Figure 5. The average time (mean  $\pm$  SE) spent on each test question in relation to the degree of engagement with the tests at institution using tests in both formative and summative tasks. Peep & sample = incomplete tests; test, review & toedip = between 1 & 2 tests completed; retest = 2-27 tests completed.**

Their extra efforts did not yield any obvious performance advantage in the final MCQ exam in the topic covered by the feedback-enriched online test (Figure 3), nor could the extra time they spent on each question be accounted for by language background.

At the two institutions where the test was freely available approximately 40% of students logged on to use it outside of the hours of 9am to 5pm. That the limited times of access to summative testing had an impact on student performance was indicated by the behaviour of a group of externally-enrolled students of the institution offering only the compulsory summative task who, being residents of the metropolitan area, were required to attend on-campus for testing. Despite above-average grades on other aspects of the course, not one of these students presented themselves, and thereby forfeited almost one quarter of their course marks

## Discussion

Our study indicates that there is no sign yet of students turning away from paid employment back to full-time engagement in their university studies. Levels of paid employment in our cohort were 10% to 20% above those described by Krause (2005) only two years ago, with similar or higher proportions working outside the university for more than 20 hours each week. The need for the flexible delivery of course materials and assessment which can be provided online cannot be said to have diminished.

It was interesting to find quantitative confirmation of Charlesworth and Vician's (2003) anecdotal observation that, when left to use online tests as they wish,

students like to be able to take breaks. This behaviour, comments provided in an online survey and the failure of younger students to persevere with completion of a paper-based questionnaire provided multiple lines of evidence for the need not only of flexibility of time constraints in online testing, but for the restructuring of tests to maintain engagement, and for exercises to develop the stamina and concentration of our younger students. Since the completion of this study we have restructured the feedback-enriched online tests presented at the institution with 75% of its first year enrolment between 16 and 18 years of age so that only 10 questions are presented in each test. Since these 10 questions are drawn randomly from the same database as served the larger tests, the number of different tests available to a student making multiple attempts has risen considerably, and the rate of repetition of testing risen. The decay over a few weeks of the advantage gained by using the online tests we revealed points to the need for repeated access to online tests for consolidation of learning and the superiority of the effect gained by repetition to its effectiveness. It was interesting to note that a number of students returned to the online tests after their final examinations. While acknowledging the advantages of this approach, we still plan to trial the gradual increase in the number of questions in tests made available across the semester. We have some evidence, in the form of comments such as

“could have figured out the answers by reading the question again and thinking about how it was worded”

that repeated exposure to explanations of right and wrong answers may be encouraging more than a cursory reading of questions, but have yet to see any evidence that this translates into improved long-term learning.

We found considerable evidence that we are not realizing the potential for flexible course delivery offered by the online learning environment. Like Volery and Lord (2000) we found that when left to their own devices students take full advantage of the flexibility offered by online course activities and log on at all hours of the day and night. We do not offer them the opportunity to take summative online assessments in any location but secure labs on campus, and at any times other than regular ‘business hours’, however. That the mismatch between preferred and imposed times of access has an impact on students ability to complete our courses successfully was indicated by the failure of a group of high quality externally-enrolled students to access significant components of their summative assessment in the course at all. Their pattern of enrolment is most commonly encountered amongst students in full-time work, but is also employed by women with heavy family commitments. As Stewart (2004) says, it is a social responsibility of education to provide, as far as possible a ‘level playing field’ with respect to access for students with diverse needs.

The difficulty in realizing the ideal set out by Stewart lies with the issue of security and validity of assessment. Rowe (2004) argues that accurate assessment, including online assessment, is essential to the survival of educational institutions, for it validates student knowledge as certified by degrees and diplomas, that if an institution claims to provide this service, they

must prove to society that they do. Noting that draconian measures to reduce cheating diminish trust, and that people who feel more “distant” cheat, Rowe clearly places us in the position of finding a way of minimizing the impact of teaching while increasing the accommodation of student needs. The chief issue when questions and answers are available before the summative test becomes that of impersonation of the student who is not present and under direct observation. Possible approaches we have considered to the issue include extending the physical access to and supervision of secure computer labs to 24 hours a day (a costly option), decreasing the value of each summative test to the point where it is not worth cheating (but this does disadvantage students with special needs), instigating sectional pass requirements so that cheating on online tests cannot lead to a pass in the unit (but this raises moral dilemmas and invites student appeals against assessment) and finding a computer-based means of verification of identity (does anyone know of one?).

We shared with Morris et al (2005) the experience of having approximately one third of our students enrolled throughout the semester, sharing the same opportunities as the others and yet failing to engage with the online material available. We were able to show the detrimental effect of this lack of engagement upon their achievements in the course, but are no more able than Morris et al to see how to motivate them to sample what we have on offer. Only 2% of students who actually investigated our online tests actually turned away from them without any real attempt at the tests.

## **Conclusions**

The patterns of use of online tests have revealed two ways in which we do not appear to be best serving the needs of our first year students. We do not appear to offer sufficient flexibility of access to online summative assessments, and we press students to complete most of these tasks when they are already fatigued.

The question of how to ensure the security and validity of online testing while increasing real flexibility of access remains unresolved for us. There is no sign yet of students turning back to full-time engagement in their university studies. Levels of participation in paid employment by first year students at Australian universities have increased steadily from 1999 through 2001 to 2004 (Krause et al., 2005) and, now, 2006. We need to address this issue, as well as those of building up the abilities of young students to concentrate on academic work, and engaging that one third of the student body not managing to find their own way to effective learning opportunities.



## References

Charlesworth, P. and C. Vician. 2003. *Leveraging technology for chemical sciences education: an early assessment of WebCT usage in first-year chemistry courses*. Chemical Education Research 80(11): 1333 - 1337.

Curtis, S. and N. Shani. 2002. *The effect of taking paid employment during term-time on students' academic studies*. Journal of Further and Higher Education 26(2): 129 - 138.

GenStat version 9.1.0.147, 2006 Lawes Agricultural Trust, Rotherham, Hertfordshire, England

Krause, K., Hartley, R., James, R., & McInnis, C. (2005). *The First Year Experience in Australian Universities: Findings from a decade of national studies*. Canberra: DEST. [Available online: <http://www.cshe.unimelb.edu.au>]

Morris, L. V., C. Finnegan and S. Wu. 2005. *Tracking student behavior, persistence, and achievement in online courses*. Internet and Higher Education 8: 221 - 231.

Rowe, N. C. 2004. *Cheating in online student assessment: beyond plagiarism*. Online Journal of Distance Learning Administration. v7, n2, p1 – 10

Stewart, B. L. 2004. *Online learning: a strategy for social responsibility in educational access*. Internet and Higher Education 7: 299 - 310.

Volery, T. and D. Lord 2000. *Critical Success Factors in Online Education*. International Journal of Educational Management. v14 n5 p216-23.



# **KEY FACTORS FOR EFFECTIVE ORGANISATION OF E-ASSESSMENT**

**Cornelia Ruedel, Denise Whitelock and  
Don Mackenzie**



# Key Factors for Effective Organisation of E-Assessment

Cornelia Ruedel <sup>1</sup>, Denise Whitelock <sup>2</sup> and Don Mackenzie <sup>3</sup>  
<sup>1</sup> University of Zurich, <sup>2</sup> Open University, <sup>3</sup> University of Derby

[Cornelia.Ruedel@access.uzh.ch](mailto:Cornelia.Ruedel@access.uzh.ch)

## Abstract

The benefits of e-assessment are widely documented (Bull and McKenna 2004). However, instances of good practice have not been systematically reported. Recognising and acknowledging this gap in the research, the JISC Organisational Committee has funded a number of projects on e-assessment practice: 'E-Assessment Glossary', 'The Roadmap to E-Assessment' together with a set of case studies of innovative and effective practice.

This paper is based on the findings of the JISC Case Study Project "The innovative and effective use of E-Assessment". Members of the project team conducted over 90 interviews with teaching staff, senior management, developers and students to showcase all aspects of e-assessment. The project offered a unique opportunity to observe different organisational structures and gain inside-information about the effectiveness of a number of different applications. The 17 case studies and their follow-up surveys have been studied to identify the facilitating factors for the introduction of e-assessment and the organisational structures supporting e-assessment have also been investigated. The focus of this analysis was to study the different organisational structures and to identify patterns herein.

We suggest that the key characteristics for the typology are the position of the e-assessment within the organisational structure and the support from the senior management. Three types of organisational structures are identified by the study, which support innovative practice. These are the Central Team, the Faculty based Team and the Departmental Champion.

The Central Team offers e-assessment support and, in some cases, production services to all academics on a university-wide basis whilst the Faculty Based Team provides a more limited discipline-related service. The Departmental Champion usually implements e-assessment within his/her specific discipline and may be an early adopter or have a special interest in this area.

## Introduction

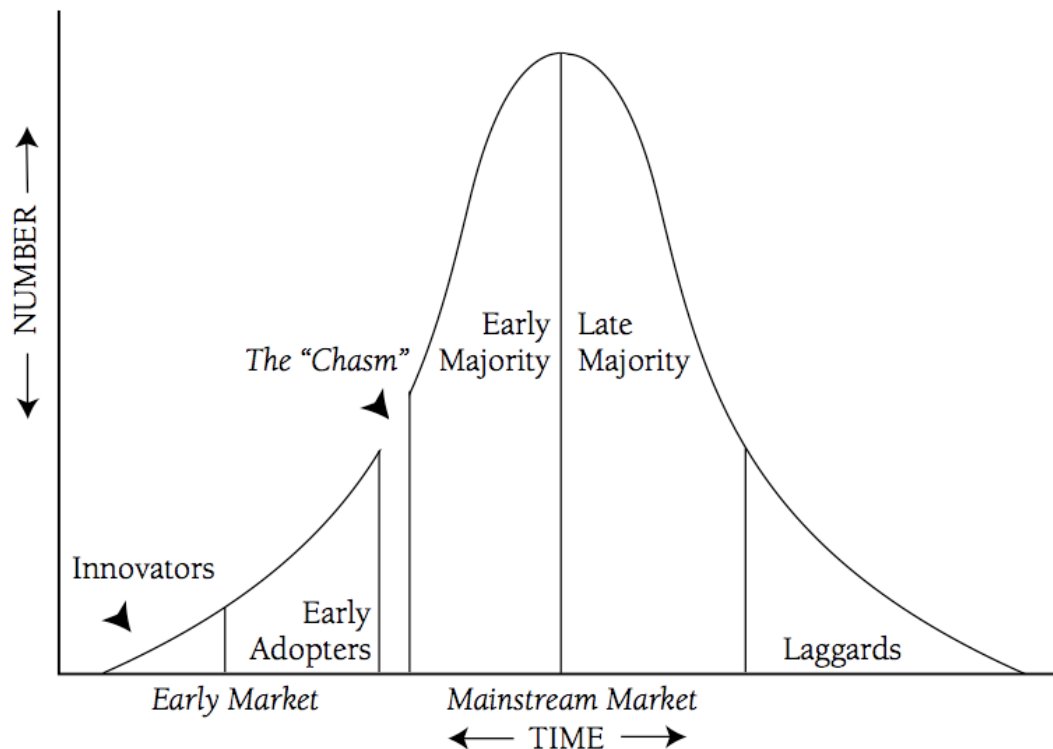
In this paper we investigate the key factors for effective organisation of e-assessment using the data collected from the JISC Case Study Project. Over 90 semi-structured interviews were conducted with practitioners, support staff, senior managers and students. During the site visits, it was observed that different institutions had diverse organisational structures in place to manage the implementation of e-assessment. This gave rise to the question of how might these organisational differences impact upon the effectiveness of promoting e-Assessment. White (2006) raises similar concerns with respect to the adoption and integration of any new technology within a given organisational structure.

## Background

The factors underlying the relatively slow and small-scale take up of e-assessment within higher education merits some investigation. A possible explanation can be found if the introduction of e-assessment is compared with the introduction of e-Learning or with the uptake of innovations in general.

For example, Warburton & Conole (2005) used the Diffusion Theory from Rogers (2003) to model the uptake of e-assessment. Rogers (2003) defines

*“An innovation is an idea, practice, or object perceived as new by an individual”*. According to Roger (1968) several variables influence the adoption of new ideas, these are: *“The situation, the personality of the adopter, the social and economic status of the adopter, the lines of communication used and the innovation itself”*. To help to understand the adoption as a process Rogers (2003) categorized the adopters into five groups using the time of the adoption as measurement. The five types of users are: Innovators, Early Adopters, Early Majority, the Late Majority and the Laggards. Geoghegan (1994) identified a 'chasm' between the early adopters and the early majority (Figure 1).



**Figure 1. Chasm between the early market and the mainstream market (Gray 1997)**

To understand this chasm it is important to understand the contrasting views and attitudes of the different types of users

<b>Early adopters</b>	<b>Early Majority</b>
Favour revolutionary change	Favour evolutionary change
Visionary	Pragmatic
Project oriented	Process oriented
Risk takers	Risk averse
Willing to experiment	Want proven practices
Generally self-sufficient	May need significant support
Horizontally connected	Vertically connected

The Early adopters want to be involved in the development of new ideas and are not afraid of failure, while the Early Majority grouping favours a more process oriented approach and wants to avoid taking risks. Therefore, these two types require different organisational support and support structures.

Geoghegan (1994) analysed the question of why information technology was not more deeply integrated into the curriculum. Several factors were identified: a shortage of equipment and facilities on campus, institutional support, unrealistic expectation of the development, use and dissemination and what he called the “Human factor”.

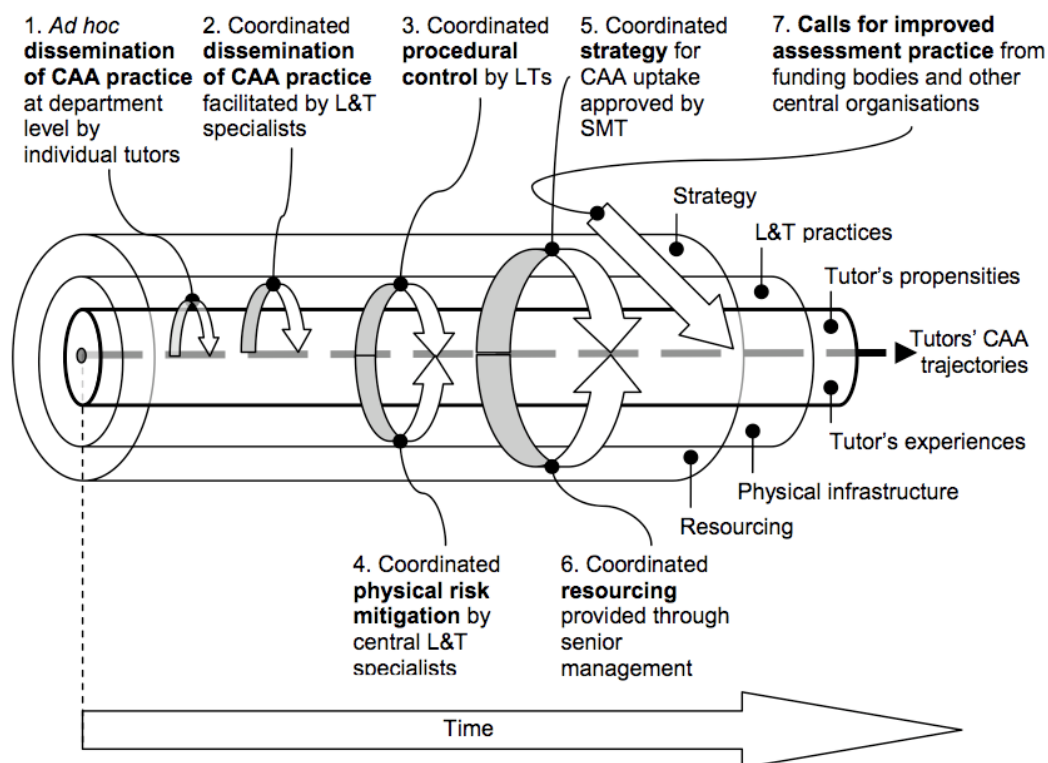
The Human factor is important in order to understand how faculties and departments interact with each other. Academics are often not at the same stage of awareness or knowledge development with respect to e-assessment

as their peers. Therefore, different support structures must be offered at different stages. Hagner (2001) introduced a classification of four types of academics regarding the adoption of innovation in an educational setting. The four types are "Entrepreneurs", "Risk Aversives", "Reward Seekers" and "Reluctants". The "First wave" of adopters or 'Entrepreneurs' are the first to adopt a new idea. They have appropriate resources either within their department/faculty or from an institutional level. The Entrepreneurs share a strong interest in improving the quality of teaching and learning and have confidence in their own expertise in order to carry a new initiative forward. On the other hand, "Second wave" users have a greater fear of the unknown or failure. They require a more persuasive and user-friendly type of support to change their well established way of teaching. "Reward Seekers" however, adopt new technology if they see a clear benefit for their career. "Reluctants" firmly "believe that traditional models of teaching are superior" (Hagner 2003).

Furthermore, the uptake on e-Learning can be taken as an example for institutional change. Within the learning technology literature, there are various descriptions of drivers and success factors. Lisewski (2004) noted, "Implementation studies of learning technology have tended to display unsophisticated perspectives on the nature of the organizational culture". They mainly concentrated on having a vision, strategic planning technical infrastructure and a strong leadership. McCartan and Hare (1996) identified four factors for change: senior management support, staff development, central services and funding opportunities. The 4-E Model was introduced by Collis and Moonen (2002) who identified the environment, educational effectiveness, ease of use and engagement as the most salient variables in their framework. Liweski (2004) too recognised a number of other factors such as 'time and space' for the innovation, effective communication at all levels, highlighting the operational aspects, staff development and a clear understanding of the requirements. Although the organisational aspect was mentioned, it was not addressed in more detail.

Walker, Adamson & Parsons (2004) did acknowledge the organisational aspects to the adoption of new technologies and recognised the presence of central support as one part of six key components of the successful delivery of e-assessment. The other five components included quality software, quality hardware, clear policies and procedures, integration within the learning system and staff education. Warburton (2006) noted that the strategic support and centralised organisational facilities are particularly evident in new universities. Existing good practice is shown as an institutional validation and as a direct impact in the uptake. A further commitment from the institutions can be seen as strengthening the physical infrastructure and secure funding. Warburton developed a concentric shell model of the CAA uptake (Figure 2) with the conditions, interactions and consequences. The conditions are divided into strategic cultural, infrastructure cultural, tutor cultural, tutor operational and infrastructure operational. Furthermore, he describes the principle mechanism driving the CAA uptake as sevenfold. The seven mechanisms are modelled upon a timeline with the starting point of ad-hoc dissemination of CAA practice at department level. The next step is the coordinated dissemination facilitated by Learning & Teaching specialists.





**Figure 2. Warburton's (2006) Concentric Model of principle mechanism driving CAA uptake**

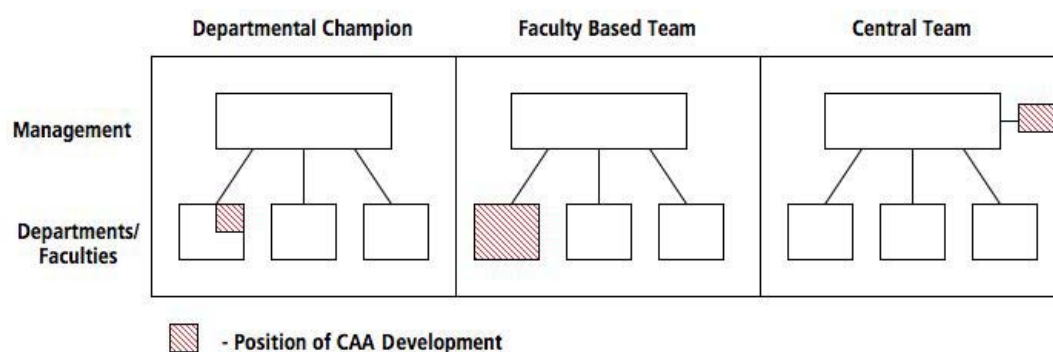
Although the model is comprehensive and explicit, the question whether the organisational structure influences the uptake on e-assessment is not raised. In Warburton's model, there seems to be no step between the development by individual tutors and the co-ordinated practice from Learning & Teaching specialists. From the observations of the e-Assessment Case Study project, there is a step in between the developments from individual tutors and coordinated practice on a departmental level as described below.

## Typology

The e-Assessment Case Study project offered the opportunity to investigate the different support structures for e-assessment in a wide variety of educational settings. The site visits, the interviews with practitioners, support staff and senior management gave a unique insight into how effective the organisational structures were and which approach works best under which conditions. The findings from the follow-up survey (Whitelock 2005a) were the basis for this categorization. The key factors which were salient to all the cases studied were the position of the e-assessment support-unit within the organisational structure and how the unit's work was embedded within the institution's e-learning strategy. Other important factors include the support from the senior management and the funding available for implementation. This paper identifies three types of organisational structures that have

resulted in innovative practice for e-assessment, these being the Central Team, the Faculty based Team and the Departmental Champion (Figure 3).

The Central Team is not attached to any department and offers its services independently to all departments or faculties. However, the Faculty-based teams are attached to only one department and the services are only available to their staff. The Departmental Champion is independent from the central services and only 1-3 tutors make use of e-assessment.



**Figure 3. Organisational support structures**

### **Departmental Champion**

The Departmental Champion is well established within the faculty or the department. The drive for any given implementation of e-assessment was to improve student learning and assessment. This group of implementers falls clearly into Rogers' category of "Innovators" and has quite a long history of development. The findings from their projects are usually well documented and disseminated nationally, although the use of e-assessment across the University is often minimal. In many cases, the purpose of the development is to demonstrate the capabilities of e-assessment and may be seen as a feasibility study. The security issues for the e-assessments are well addressed and the delivery is through a closed-network or on paper as an OMR. The projects are tailored to a specific need either pedagogical or technical. However, they are too specialised to enter the mainstream of the university's assessment strategy and often the funding for the development of a particular type of assessment comes from outside the university. The Project Team identified Departmental Champions at UCL, University of Glamorgan, University of Surrey, University of Cardiff and others undertaking innovative work in e-assessment.

### **Faculty based Team**

The Faculty-based Team centres on an enthusiastic circle of academics. It secures project funding at both the departmental level and from external sources. The e-assessment system may be commercial or developed in-house and is supported by a dedicated developer/academic or a team of developers. The team is led by an academic who inspires the pedagogical

and technical development. However, the infrastructure to facilitate e-assessment for large student groups may not be available and the students usually use the computer rooms of the faculty/department in order to sit the examinations/tests. The in-house e-assessment system may be developed up to a commercial level and all security issues of the delivery are addressed. Heriot-Watt University is an example for a Faculty based Team which has a great reputation and a long history of expertise.

The development of feedback rich formative e-assessment can be one of the features that has been particularly extended. The in-house e-assessment system or use of e-assessment in general may be part of a nation-wide project initiative and is disseminated nationally and internationally. For example, the team at Birkbeck College built on their departmental work to attract external funding for the FDTL4 – OLAAF (Online Assessment And Feedback) project that has brought together a number of faculty and departmental champion initiatives across a range of institutions. Senior management may or may not build on this approach within the participating institutions to create a university-wide initiative. Therefore, the impact of the project can still be limited to the department or faculty despite its inter-institutional success.

### **Central Team**

The Central Team develops, supports and coordinates the e-assessment activities university-wide. Commercial software (e.g. Questionmark Perception) is often installed to facilitate e-assessment, although the TRIADS system from the University of Derby is used successfully for the university-wide delivery of summative assessments. E-assessment applications are well integrated into the VLE and university processes and can be accessed anytime and anywhere in some cases. Students use it for summative or formative assessment and are aware of the benefits.

The Central Unit may act as a facilitator for individual academics wishing to deliver e-assessments or it may go further and offer a complete consultancy, production, delivery and results reporting service. Mackenzie (2005) has outlined the relative benefits of the latter in terms of quality assurance of summative assessments when compared to a devolved tutor development approach.

The senior management of the educational institutions have invested in the infrastructure for the delivery of e-assessments. Computer laboratories are available for up to 200 students with separate entrances and exits and may be equipped to conform with the guidelines laid out in the BS7988 / ISO/IEC DIS 23988 'Code of practice for the use of information technology (IT) in the delivery of assessments'.

IT services and the central unit work closely together and have published procedures and guidelines to clearly identify all the tasks for the different

teams and services. The central unit may be integrated into the Centre of Teaching and Learning/Educational Development.

There are a number of communication channels, which disseminate the innovation 'E-Assessment' to the academics including a staff development programme. The most effective network seems to be where faculty-based E-Learning Coordinators, which is use for example at the University of Southampton and Loughborough University. The Coordinators inform the teaching staff about e-Learning in general and the possible uses of e-assessment in particular. This e-learning communication network seems to work effectively and even reaches tutors beyond the early adopters.

An alternative to e-Learning Coordinators is the use of academics in the role of e-Champions, which is used at the University of Derby in addition to Teaching Fellows with responsibility for e-learning. According to Rogers (2003), champions should be "charismatic" individuals who throw their weight behind the innovation". Information is more widely spread if it comes from a trusted source like a fellow academic. Drawbacks are that the workload of academics nowadays has increased dramatically and to sustain this type of initiative the individual champion needs to have enough time and energy for the full benefits to be realised by the parent institution. Staff development too needs to be offered on a number of various levels to cater for the different skills of the tutors. It is vital that the academics can choose the type of training that supports their own requirements. The most frequently used form of staff development has been the workshop or seminar while one-to-one consultations have been offered for more specific problems. The provision of structured online courses for tutors has been found to be very successful, for example, the 'Assessing Online' module at the University of Dundee (Walker 2004).

## **Discussion and Conclusions**

When reviewing the development of e-assessment via the three different organisational structures noted above there seems to be a correlation between the provision of a 'central' support team and the effective adoption of e-assessment.

The key factors of effective support are:

- The appropriate position, status and role of the e-Assessment Unit
- Effective communication channels incl. staff development
- Availability of respected and experienced 'champions'
- Support from senior management

The positioning of a Central E-Assessment Unit so that it is accessible to all academics on a university wide basis seems to be the key for successful delivery and dissemination because it demonstrates the commitment of the senior management to support e-assessment and demonstrates their

confidence in its effectiveness. Equally important is that the Unit works closely with the technical units (IT) and has the input from the pedagogical centre to provide integrated support.

It is helpful if formative e-assessment can be embedded into the virtual learning environment (VLE) and into the IT system available to academics on their desktop. Often the introduction of a VLE can be seen as a catalyst for a university-wide implementation of e-assessment. Embedding e-assessment activities into the VLE may help to 'kick-start' wider implementation, and even though the native VLE system may provide little more than 'quiz' functionality, it can lead to the adoption of more sophisticated systems and development of summative assessment as experience is gained.

The communication channels used to introduce and establish e-assessment as a valid tool plays a vital role. Staff Development programmes represent the traditional approach for training tutors in the effective use of e-assessment. Key to this approach is the availability of high quality, subject-specific exemplars. Academics new to this form of assessment often find it difficult to relate to examples outside their own discipline.

New methods such as the adoption of E-Learning-Coordinators or E-Learning Champions are being explored in many institutions to reach even more staff.

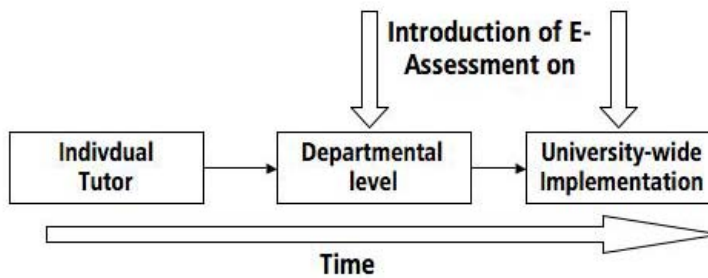
The support from senior management is significant for the delivery of e-assessment. Investment into the development of the e-assessment and into the infrastructure necessary to deliver it demonstrate the commitment of the management to implement the E-Learning Strategy and helps to enhance the status of early adopters and champions.

The three types of organisation outlined above could be seen merely as a classification system. On the other hand they may reflect stages in the natural evolution to more widespread adoption of e-Assessment within an institution outlined below and in Figure 4.

Stage 1: Enthusiastic academics develop/use an e-assessment tool which is used to deliver assessments to his/her students in the first instance. The findings of the pilot project are disseminated within the department/externally and fellow academics use the system to deliver more assessments.

Stage 2: Further funding from the department or external bodies facilitates enhanced software development or more sophisticated e-assessments. More widespread dissemination, both internally and externally can be used to validate the academic credibility of the systems or assessments and to bring the developments to the attention of senior management within the institution.

Stage 3: The senior management of the institution acknowledges the development and initiates a central support unit to establish e-assessment as a credible and valid tool for learning and examination and provides an academic support infrastructure that encourages the development of e-assessment embedded in e-learning.



**Figure 4. Organisational support structures**

The outcomes of the Project indicate that typical timescales for this evolution from early adopter to relatively mature and widespread implementation have been of the order of ten years or more in those institutions where e-assessment is currently well developed.

Good internal communication and dissemination of grass roots developments to the highest management level within the institution is key to the successful movement between these phases. On the other hand, top-down imposition of e-assessment methods without sufficient support or pedagogically sound exemplars from well respected members of staff has the potential to promote resistance and slow development.

Recognition of the stages outlined above should help institutions to identify the actions necessary to progress through the organisational and infrastructural barriers between them to a more widespread adoption of appropriate application of many types of innovation.

As observed in one or two cases studied during the Project, progression beyond the three stages outlined above may lead to the commercialisation of e-assessment software, bespoke e-assessment development (University of Derby) or e-assessment training (University of Dundee) that is capable of generating external income for institutions that are prepared to grasp the nettle and invest in the appropriate staff and infrastructure.

## References

Baggott, G & Rayne, R, (2007) OLAAF Online Assessment and Feedback Project Web Site. <http://www.bbk.ac.uk/olaaf/>

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

Collis, B. and Moonen, J. (2002) Flexible learning in a digital world, *Open Learning*, 17, pp. 217-230.

Geoghegan, W. H. (1994) What Ever happened to Instructional Technology. In: Bapna, B, Emdad, A. and Zaveri, J. (eds.) *Proceedings of the 22<sup>nd</sup> Annual Conference of the International Business Schools Computing Association*. Baltimore: International Business Schools Computing Association

Gray, P. (1997) Viewing Assessment as an Innovation: Leadership and change process. *New Directions for Higher Education*, 25 (4), pp. 5-15

Hagner, P. (2001) Interesting practices and best systems in faculty engagement and support. Presentation at NLII Focus Session in Feb. 2000, Seattle, USA

Liweski, B. (2004) Implementing a learning technology strategy: top-down strategy meets bottom-up culture. *ALT-J Research in Learning Technology*, Vol. 12 (2), pp. 175-188

Mackenzie, D.M. (2005) Online Assessment: quality production and delivery for higher education. Keynote Address in Enhancing Practice, Assessment Workshop Series No. 5 in Reflections on Assessment, Volume II. pp22-29, Gloucester: Quality Assurance Agency for Higher Education

McCartan, A. and Hare, C. (1996) Effective institutional change: the impact of some strategic issues in the integrative use of IT in teaching and learning. *ALT-J Research in Learning Technology*, Vol. 4, 21-28.

Rogers, E.M. (1968) The Communications of Innovations in a Complex Institution. *Educational Record*, pp. 67-77

Roger, E.M. (2003) *Diffusion of Innovations*. New York: Free Press

Stephens D, Bull J, Wade W. (1998) Computer-assisted assessment: suggested guidelines for an institutional strategy. *Assessment & Evaluation in Higher Education*, 23 (3), pp. 283–294.

Technical Committee IST/43, British Standards Institute, 2002, BS7988, A Code of practice for the use of information technology (IT) for the delivery of assessments.

Walker, D. Adamson, M. Parsons, R. (2004) Staff Education – Learning about Online Assessment, Online. In: Danson, M., ed., *Proceedings of 8th International CAA Conference*. Loughborough, University of Loughborough.

Warburton, W. and Conole, G. (2005) Whither E-Assessment? In: Danson, M., ed., *Proceedings of 9th International CAA Conference*. Loughborough, University of Loughborough.

Warburton, W. (2006) Quick win or slow burn? Modelling UK HE CAA uptake., In: Danson, M., ed., *Proceedings of 10th International CAA Conference*. Loughborough, University of Loughborough.

White, S. A. (2006) Critical Success Factors for Institutional Change: Some Organizational Perspectives. In *Proceedings of Critical Success Factors for Institutional Change, a workshop of the European conference of Digital Libraries, (ECDL'06)*, pp. 75-89, Alicante, Spain. Davis, H. C. and Eales, S., Eds.

Whitelock, D., Mackenzie, D., Whitehouse, C., Ruedel, C. and Rae, S. (2006). Identifying Innovative and Effective Practice in e-Assessment, JISC Report



# **QUALITY ASPECTS OF OPEN SOURCE TESTING TOOLS**

**Friedrich Scheuermann  
Ângela Guimarães Pereira**



# Quality Aspects of Open Source Testing Tools

Friedrich Scheuermann and Ângela Guimarães Pereira  
European Commission - Joint Research Centre, IPSC  
Knowledge Assessment Methodologies (KAM)  
TP 361, Via Enrico Fermi 1 / 21020 Ispra, Italy  
[friedrich.scheuermann@jrc.it](mailto:friedrich.scheuermann@jrc.it), [angela.pereira@jrc.it](mailto:angela.pereira@jrc.it)

## Abstract

This paper presents work in progress concerning the definition of quality criteria for open source computer based assessment, namely platforms for the assessment of skills. The research approach undertaken so far is based on literature reviews and expert interviews which contributed to identify a number of software applications, platforms and tools being currently reviewed according to a pre-defined matrix of descriptive and normative criteria. The results of the evaluation activities will feed the setting-up of a protocol for quality assurance of e-assessment platforms in skills assessment contexts.

## Background

In 2006 the European Parliament and the Council of Europe have passed recommendations on key competences for lifelong learning and the use of a common reference tool to observe and promote progress in terms of the achievement of goals formulated in “*Lisbon strategy*” in March 2000 (revised in 2006, see <http://ec.europa.eu/growthandjobs/>) and its follow-up declarations in selected areas (Communication in the mother tongue, communication in foreign languages, mathematical competence and basic competences in science and technology, digital competence, learning to learn, social and civic competences, sense of initiative and entrepreneurship, and cultural awareness and expression) (European Parliament and Council of Europe, 2006). Indicators for the identification of such skills are now needed, as well as instruments for carrying out large-scale assessments in Europe. In this context it is hoped that electronic testing could improve the effectiveness of the needed assessments, i.e. improve identification of skills, and their efficiency, by reducing costs of the whole operation (financial efforts, human resources etc.).

This paper describes developments within a project on e-assessment quality assessment whose overall aim is the development of quality criteria to assess e-assessment platforms and draft recommendations for such systems in contexts of skills assessment (including desirable architectures, required competencies, interoperability requirements, etc.).

In the remainder of this paper we will describe the methodology and preliminary results of a review of practice on computer based assessment, focussing on open source software applications, though including commercial options. This review's results are the basis for developing such protocol.

## Research design

The research approach is framed by the need to assess skills of population groups in Europe at a large scale and to achieve accurate and comparable results for further benchmarking. Therefore, emphasis is given to tools for **diagnostic assessment and objective measurement** as the basis of research activities on *e-assessment*.

The following research questions were formulated for further orientation of the work:

1. **Potentials:** What are the potentials of testing software in relation to existing instruments for measurement? What are the implications for policy and lifelong learning?
2. **Requirements:** What types of platforms are needed in order to carry out large-scale testing in a very heterogeneous European environment which is also characterised by different infrastructures, possibilities and needs in terms of technology? What are the requirements to be respected, functionalities and features need to be taken into account for delivery?
3. **Open Source:** What is the specific added value of open source software in the context of assessment? What are the characteristics? How is it being implemented? and what are existing relevant experiences?
4. **Quality:** What are the quality dimensions to be taken into account? Which criteria can be applied for the definition of quality in open source platforms and the delivery of tests?

These questions are probed into the differential experiences of actors, such as policy-makers, test developers, test takers and test administrators, being derived from literature reviews and interviews.

Furthermore, an in-depth evaluation of a selected choice of platforms drawn from a vast range of tools identified in Internet sites and literature, using a pre-defined matrix is carried out. The evaluation is based on a mix of inspection and test methods applied to system usability as well as taking into account different phases and stages during the broader context of the assessment process.

The results of the work will be revised in several steps through a peer-reviewed process with European expert researchers and practitioners.

## Instruments

The matrix of criteria for software evaluation was produced based on literature review and internet search. This is an on-going process, which also allows addressing the general context of testing and to identify relevant products and methodologies, as well as key actors in the field. Based on this research, an overall analysis of potentials and threads from a user's perspective (test taker, test developer, test administrator) was carried out and set into context of selected application areas, such as languages.

The evaluation matrix is composed of a set of categories derived from literature review and refined by the analysis of a selected number of randomly chosen applications. The matrix takes stock of initial work by Bergstrom et al. (2006) who have developed and applied a tool for assessment and online delivery. Apart from administrative data the adapted matrix contains assessment items, such as:

- Availability (URL, CD-ROM, Demo etc.)
- Licence/Costs (Open Source, Freeware, Commercial etc.)
- Delivery Method (Internet/Web-based, stand alone, secure site)
- Type (tool, platform, service etc.)
- General features (Specific assessment functionalities, administrative functionalities, communication etc.)
- Field of Application (context, such as Languages, personal skills assessment)
- Purpose (e.g. self-assessment, peer-assessment)
- Function (diagnostic, summative, formative)
- Target group(s) (Age, profession etc.)
- Outcomes (expected outcome of assessment activity, to which the tool is enabling)
- Item Types (MC, open questions etc.)
- Language(s)
- Standards (Is reference made to any applied standard?)
- Quality assurance (Is reference made to any specific quality assurance measure?)
- Interface/ Access Restrictions (e.g. open access, restricted access)
- Hardware/Software Requirements
- Stakes (high, medium, low)
- Assessment algorithms?

A first categorisation of products aimed at selecting platforms according to their relevance for the project. Categorisation and relevance of software is based on the degree of compliance of the platforms for the following features:

- Diagnostic testing

- Objective measurement
- Platform, covering all phases and steps of assessment
- Proctored, internet-based assessment features
- Multilingual or potential to deliver in multilingual versions
- Availability/accessibility for evaluation
- At a later stage: open source license

Finally, contextualised experiments will be carried out with a limited number of software products in order to identify and verify quality indicators, which in turn will contribute to a first version of a quality criteria checklist for e-assessment platforms. The work will be peer reviewed by experts' workshops, leading to a tuned version of such platforms.

### **Platform evaluation**

The starting point of the analysis is the expected benefit from testing measurements in general and from supportive electronic environments. Testing activities can be fully based on ICT platforms or just enhanced by ICT in addition to other forms of the assessment process (e.g. some types of "blended assessment" mixing different ways of delivery). From our revision of the existing literature it seems as though that there are almost as many criteria as there are contexts, scenarios and stages for testing. Such criteria relate to the adequateness of assessment methodologies (from a psychological/psychometrical, pedagogical perspective), technical features and specifications as well as to socio-economic reflections. However, few experiences are documented to provide a sound overall picture of the complete scope and process of effective and efficient computer-based test delivery.

A first classification of products and services aimed at separating those items into those of relevance for this project. They were classified according to the above mentioned types and then selected on the basis of availability, features provided and licences. Separation of software into open-source and commercial (including shareware, freeware etc.) types was not considered to be appropriate at this stage since we would like to keep an overview of the state-of-the-art and innovative solutions, which outlines promising potentials for future applications in skills assessment, in particular.

There exist a large number of electronic tools on the market supporting assessment activities. Such tools are offered either as

- specific functionality of (educational) platforms that enable the management of (usually multiple-choice) items together with the administration and server- or web-based delivery of tests (e.g. Moodle, <http://www.moodle.org> ),
- survey development tools (e.g. Hot Potatoes, <http://hotpot.uvic.ca>),

- tools dedicated to data collection and analysis of results/measurement (e.g. OpenSurveyPilot, <http://www.opensurveypilot.org/>)
- management tools, e.g. for documentation, reporting (including grading tools, classroom/pupil assessment administration ) (e.g. Gradebook 2.0, <http://www.winsite.com/bin/Info?2500000035898>)
- assessment platforms, covering the complete process of assessment activities (e.g. TAO, <http://www.tao.lu>), or
- assessment services (e.g. Pan Testing) covering a wide range of (tailor-made or standard) activities proposed depending on specific needs. Such services are usually offered by commercial enterprises (ASP).

So far, based on literature review and internet search, more than 460 products and services were identified which then have been explored and classified according to the categories defined earlier. As a consequence, based on the features listed earlier, a list of assessment platforms was derived, out of which 3-5 will be tested in a next step of the project.

Many tools and applications are being developed by commercial enterprises with specific services on well-focussed areas. However, availability of platforms for test delivery is limited. An example for such a platform is TAO (*Test Assisté par Ordinateur*) system (See: Plichart, Jadoul et al. 2004 and <http://www.tao.lu>) TAO is a modular platform for internet-based computer aided testing. The platform allows the management of knowledge pertaining to subjects (individuals whose competencies and knowledge may be assessed), groups of subjects, tests and items (elements of tests requiring an answer from the user). TAO is said to be a flexible and distributed system since it uses meta-data for resource description formalised through Semantic Web standard language RDF/S. In the words of the TAO authors any sort of testing in several domains, including accreditation and even surveying could usefully deploy this open source (OSS).platform. This system is still under development, although a full prototype already exists. The TAO system has not undergone major testing. Also, according to the authors it has much more potential than existing assessment platforms, being a dedicated assessment platform, the elements and properties of which, provide the link with psychometric theory (item parameters and characteristics, testing algorithms etc.) being explicitly built into TAO, but still open for relevant tailoring. The platform is in principle interoperable with other electronic applications.

Its main assets, regard the open shell concept that allows easily specific functionality to be added as a plug-in; currently it includes a variety of assessment models, as well as possibilities for having construction of items other than just multiple choice, in addition to a user friendly interface from the point of view of the test taker. However, the platform is not yet developed on industrial standards due to lack of funding.

One of the reasons to go Open Source in these types of platforms is to try to boost through a community of users further developments. This project will try

to verify this statement at a later stage. During our software review, a great deal of that what is presented under this *branding* is not corresponding to that what is commonly understood as “Open Source” in terms of the availability of open source code (see for instance the OSI, <http://www.opensource.org/>). In many cases this software is declared as “work in progress” to be published at a later stage or, as in most cases, out of date and not anymore accessible.

### **Final remarks**

Results of the analysis of selected platforms will be presented during the conference event. Furthermore, a preliminary version of quality indicators and criteria will also be presented.



## References

Bergstrom et al. (2006). Defining Online Assessment for the Adult Learning Market. In: Online Assessment and Measurement. M. Hricko and S.L. Howell. Hershey, London, Information Science Publishing: 46-47.

European Parliament and Council of Europe (2006). Recommendation of the European Parliament and of the Council on key competences for lifelong learning: 10-18

Plichart, P. et al (2004). TAO, a collaborative distributed computer-based assessment framework built on Semantic Web standards. In: International Conference on Advances in Intelligent Systems – Theory and Applications; AISTA2004. Luxembourg



# **EXAMONLINE : E-ENABLING “TRADITIONAL” HE/FE EXAMINATIONS**

**Felix Schmid, Tom Mitchell, Jacqui Whitehouse,  
Peter Broomhead**



# **Examonline: e-Enabling “Traditional” HE/FE Examinations**

Felix Schmid  
University of Birmingham  
Tel: 0121 414 5138  
f.schmid@bham.ac.uk

Tom Mitchell  
Intelligent Assessment Technologies Ltd,  
Tel: 01555 660688  
tom@intelligentassessment.com

Jacqui Whitehouse  
University of Birmingham  
Tel: 0121 414 4191  
j.whitehouse@bham.ac.uk

Peter Broomhead  
Brunel University  
Tel: 01895 265775  
peter.broomhead@brunel.ac.uk

## **Abstract**

The authors of the present paper describe ExamOnline, a system specifically developed to e-enable summative essay style examinations delivered in a Higher and Further Education setting. They also discuss the results of two live pilots, undertaken with the same group of students but with two different versions of the system. The system has been specifically designed to support existing examination processes, such as clerical level document-based authoring and distributed assessment by multiple markers. From an educational perspective, the aims are to provide a better and more relevant examination experience for increasingly computer literate student cohorts and to support effective blind marking of on-screen student responses. The authors also seek operational efficiencies in terms of paper-free streamlined administration and marking. The results of two live pilots indicate that the system achieves the objectives in terms of both the student experience and staff perception of fairness in assessment.

## **Introduction**

The vast majority of HE/FE summative examinations are not composed of atomic, closed form assessment units, such as Multiple Choice Questions (MCQ). Instead, they consist of questions requiring extended responses and essay type answers. Educators dread the marking burden such examinations impose, but they are firmly wedded to the perceived advantages of the assessment instrument in measuring a student's understanding rather than their ability to retain data. For the bulk of UK HE/FE examinations, therefore, there is little imminent likelihood of a mass migration to on-screen automatically marked tests. Nevertheless there are important drivers to move away from paper based tests and towards on-screen assessment.

Not the least of these relates to the quality and relevance of the assessment experience for students when generating extended response or essay answers without recourse to that now ubiquitous tool, a word processor. The increasingly anachronistic constraints imposed by hand written examinations (no cut-and-paste, no formatting, no spell check, etc) bring into question the very fitness for purpose of a script based examination process for an increasingly digital cohort (Prensky, 2001).

Reducing or removing the more undesirable aspects of subjective marking is another driver. Blind marking, easily implementable in an electronic system, is eminently desirable, as is removing the (possibly subconscious) influence on marking of poor (and often illegible) handwriting and, increasingly, poor spelling and grammar.

The move towards on-screen delivery and marking of “traditional” examinations is, potentially at least, problematic. The now familiar issues relating to delivering tests on-screen (Conole, 2005; Sim, Horton, 2005) are potentially magnified in a test where extensive typing is required. Moreover, existing commercial e-assessment platforms are geared towards the delivery of closed form items, automated marking, and item banking, and are not generally modelled on the classical HE/FE examination model. Creating and administering tests on these systems is often the domain of specialist learning technologists. This is in stark contrast to the existing situation with paper-based examinations, where the administration of the examination process (i.e., test paper formatting, photocopying, distribution, marks accumulation and output) is typically in the hands of academic, administrative and clerical staff. Assessment is generally carried out by individuals or teams of academics who literally ‘mark’ written responses in the answer books.

## **Background**

For the reasons mentioned in the introduction, the team managing the MSc programme in Railway Systems Engineering and Integration (RSEI) at the University of Birmingham have embarked on the production of an end-to-end on-screen examination system specifically tuned to the requirements of the

HE/FE summative examination process. They chose a small commercial supplier to adapt well tested software to cope with the new demands.

The RSEI programme is an interdisciplinary postgraduate course. The programme has a strong focus on developing individuals' railway engineering knowledge and their systems integration skills. Many participants are experienced railway engineers and managers, sponsored by their employers. The taught part of the programme in RSEI is built around eight assessed modules of 15 credits each, four supplementary modules and an integrating dissertation attracting 60 credits, all at the Masters level. An assessed module involves about 30 hours of teaching, 20 hours of tutorials, a major team exercise and some 90 hours of independent study. The assessment of learning is based on class tests, assignments and end of year examinations.

The system described in this paper has initially been used to deliver two class-test type summative examinations to MSc students enrolled on the RSEI programme, during the 2007 spring semester. The system will be rolled out on a wider basis during the remainder of this academic year and into the next. The name of the system is ExamOnline.

### **The HE / FE Examinations Model**

ExamOnline has been specifically developed to e-enable summative essay style examinations delivered in an HE/FE setting. Assumptions for the examination model are as follows.

- Examinations will generally consist of essay / extended response / short answer questions (and possibly a mixture of all three);
- Some questions will require drawings and calculations to be assessed as part of the process;
- Examinations will be invigilated through the physical presence of staff, on University / College premises;
- Examinations will require detailed human marking, often involving multiple markers.

### **User Requirements for the System**

The user requirements that were identified for the e-enabled examination system can be summarised as follows:

- It must be specifically designed to support the prevailing HE/FE examinations model, as outlined above;
- It should be simple to use. Specifically:
  - Present an intuitive and, where possible, familiar interface to students, invigilators, assessors and administrators;
  - Be web-based, with all content presented in a standard web browser, with no client software installation required.

- It should provide an efficient, streamlined examination process, specifically incorporating the following features:
  - Support secure, distributed, on-screen marking by multiple markers;
  - Support secure “offline” marking such that markers can download data to laptops for marking “as and when”, and support the subsequent synchronisation of data on upload;
  - Support simple, intuitive, document-based test authoring and results output, suitable for use by administrative / clerical staff and, in some cases, academics.

ExamOnline has initially been used to deliver two summative examinations to MSc students during the 2007 spring semester, with further summative examinations to follow shortly. An overview of the system and a description of the key design issues and features are presented in the following sections.

## The ExamOnline System

Key aspects of system design are outlined in the following sections. Planned extensions and developments are mentioned in outline only.

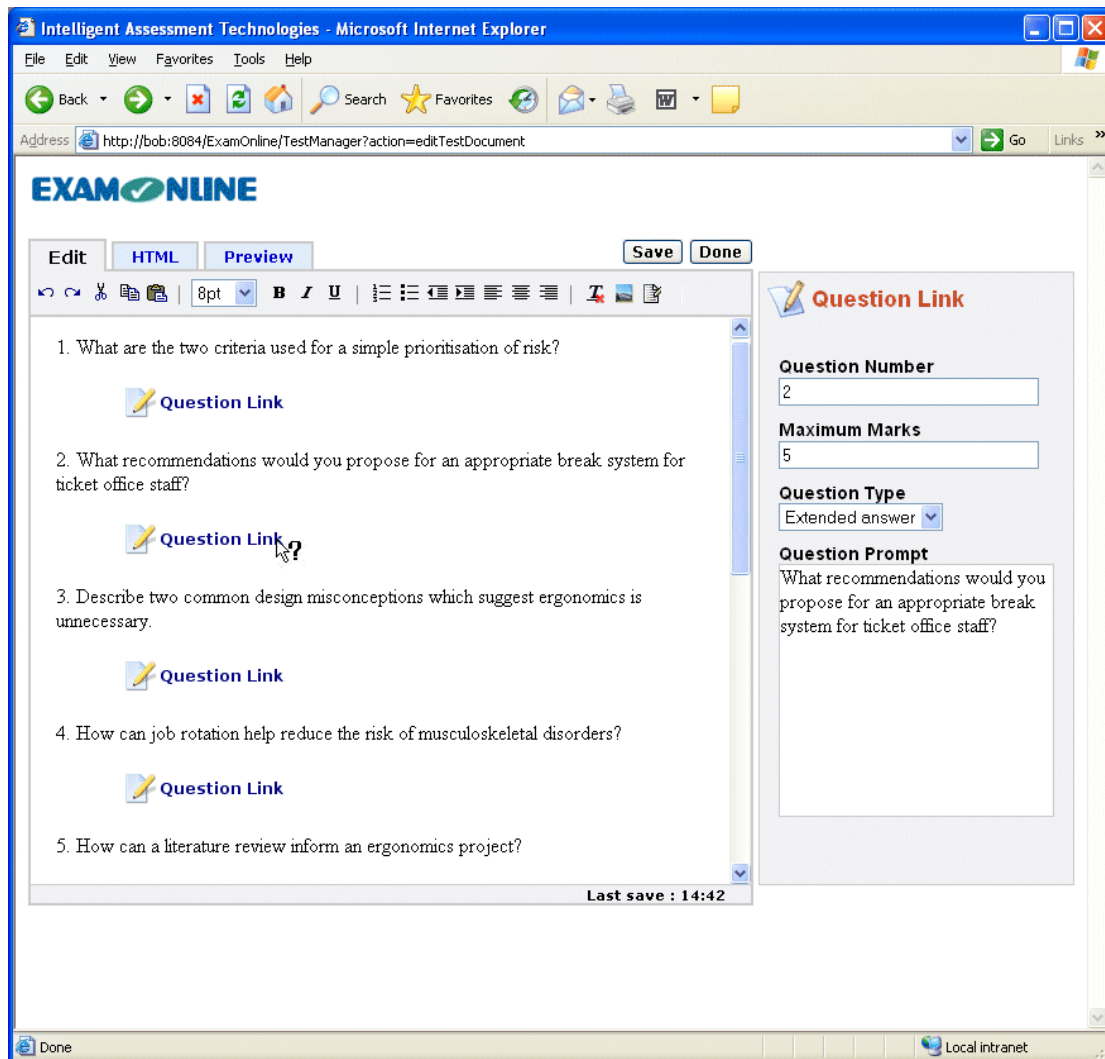
### *Test Authoring*

The system has been modelled on the existing examination development process, accepting that most examination papers at advanced level consist of text based questions with occasional graphics. A key part of this is that exam papers are developed as **documents** – they not have to be assembled from banks of items. As a matter of course, authors will routinely cut and paste (and possibly modify) existing questions from previous exam papers but, typically they do not maintain a bank of questions in the form that is familiar to proponents of computer-based testing. The tool that they will use for creating examination papers is, of course, a word processor.

For these reasons, the test creation interface for ExamOnline is just that, a simple web-based WYSIWYG (What You See Is What You Get) word processor which supports text formatting, inclusion of graphics and cut and paste from traditional word processor applications, such as Microsoft Word.

The system does not require the creation of individual items, but rather that of examination documents (test papers), into which the persons in charge of exam paper production insert ‘question links’ using the authoring interface (see Figure 1). When the paper is delivered (i.e. when the on-screen examination takes place) the students will click on these links to answer individual questions. The entire process is geared towards the existing skills of administrative staff in producing word-processed documents.

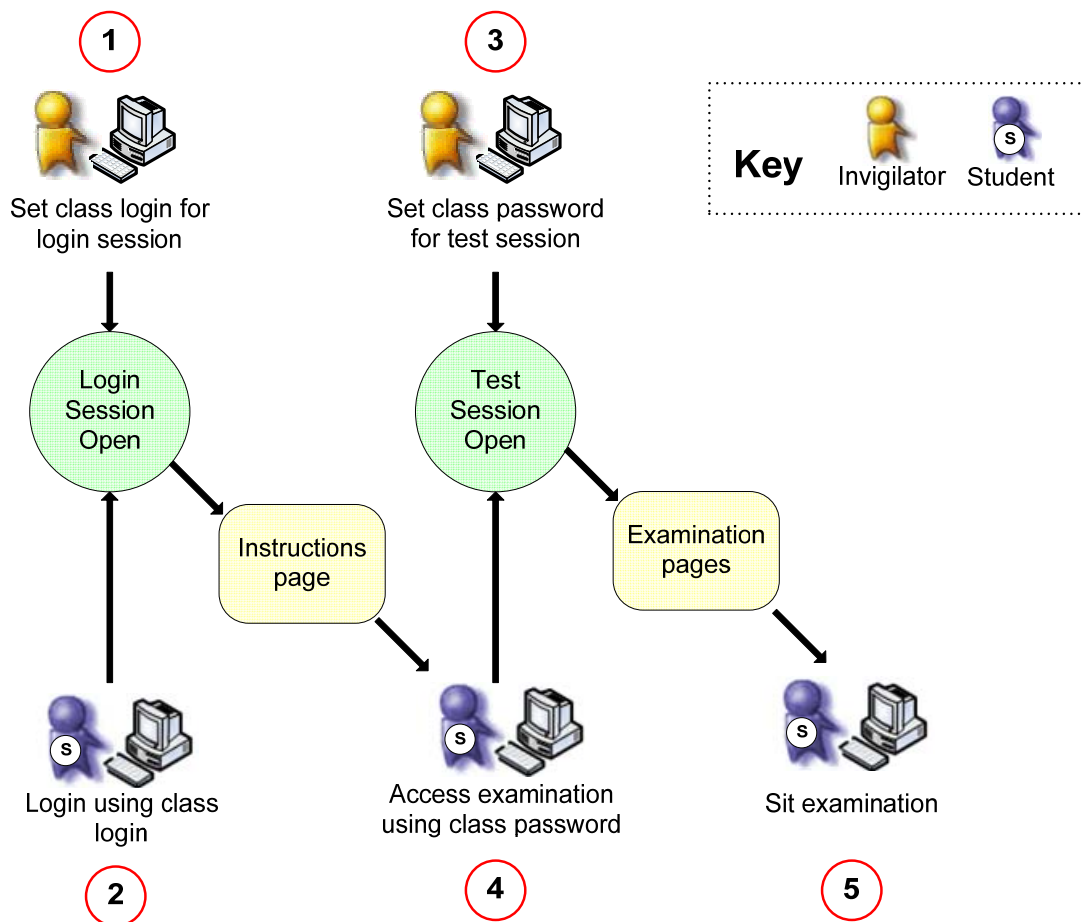




**Figure 1: Screenshot of test authoring interface**

### *Administering an Examination*

The system is based on the well-known model of an invigilated examination, with the invigilator in the same room as the students sitting the examination. The invigilator logs in to ExamOnline using a PC, or another web-enabled device, in the exam room itself, and the system provides a simple point and click interface for the invigilator to select the examination paper. The next step is to open a login session for the examination and to specify a class login. The class login is released to the students, most likely on an OHP or data projector. The candidates are then able to login to the system and to proceed as far as a holding page, which gives instructions on the test. When ready, the invigilator will open the test session proper, specifying the duration of the session and, at the same time, defines a class password for it. When this is released to the students, they can begin the examination. The diagram in Figure 2 illustrates this process.

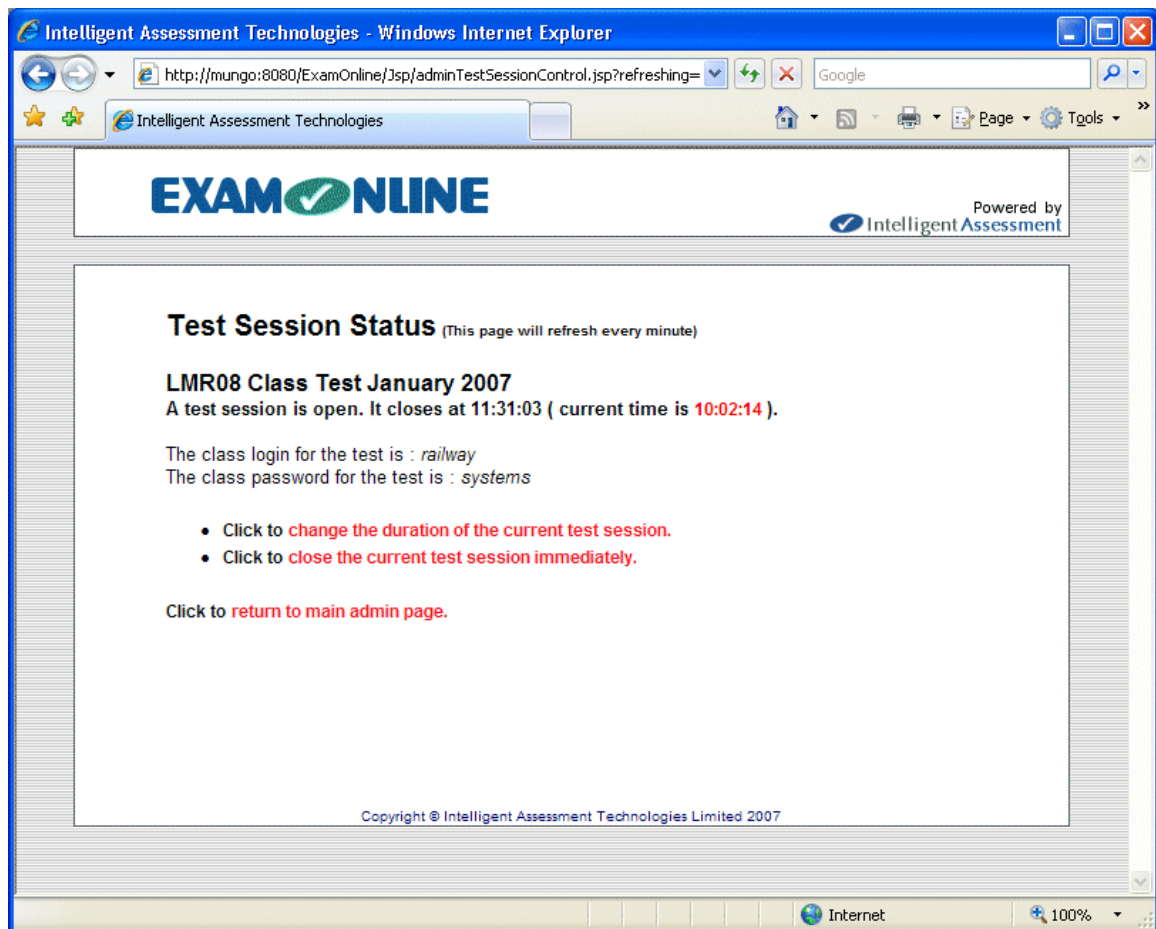


**Figure 2: ExamOnline's staged login process puts invigilators in control**

There are three important points to note:

- There is no notion of candidate registration in the system – that is, there is no need to create a list of participants who are expected to take the examination. Rather, the system simply creates a unique test session for each student who logs in to take the test. As with a normal examination, verification of student identity (e.g., by means of a matriculation card) is the duty of the invigilator;
- Accordingly, the system uses a confirmatory login process, where students are asked to re-enter key information (e.g., their matriculation number) to ensure valid input;
- Students leaving the examination room and subsequently trying to re-login to the test will not be able to access their test – a second level administration password is required for re-logins.

The screenshot in Figure 3 shows the invigilator's view of an open test session. The screen displays the login details that the candidates need to access the examination and enables the invigilator to change the duration of the session, if required.



**Figure 3: Invigilator's view of an open test session**

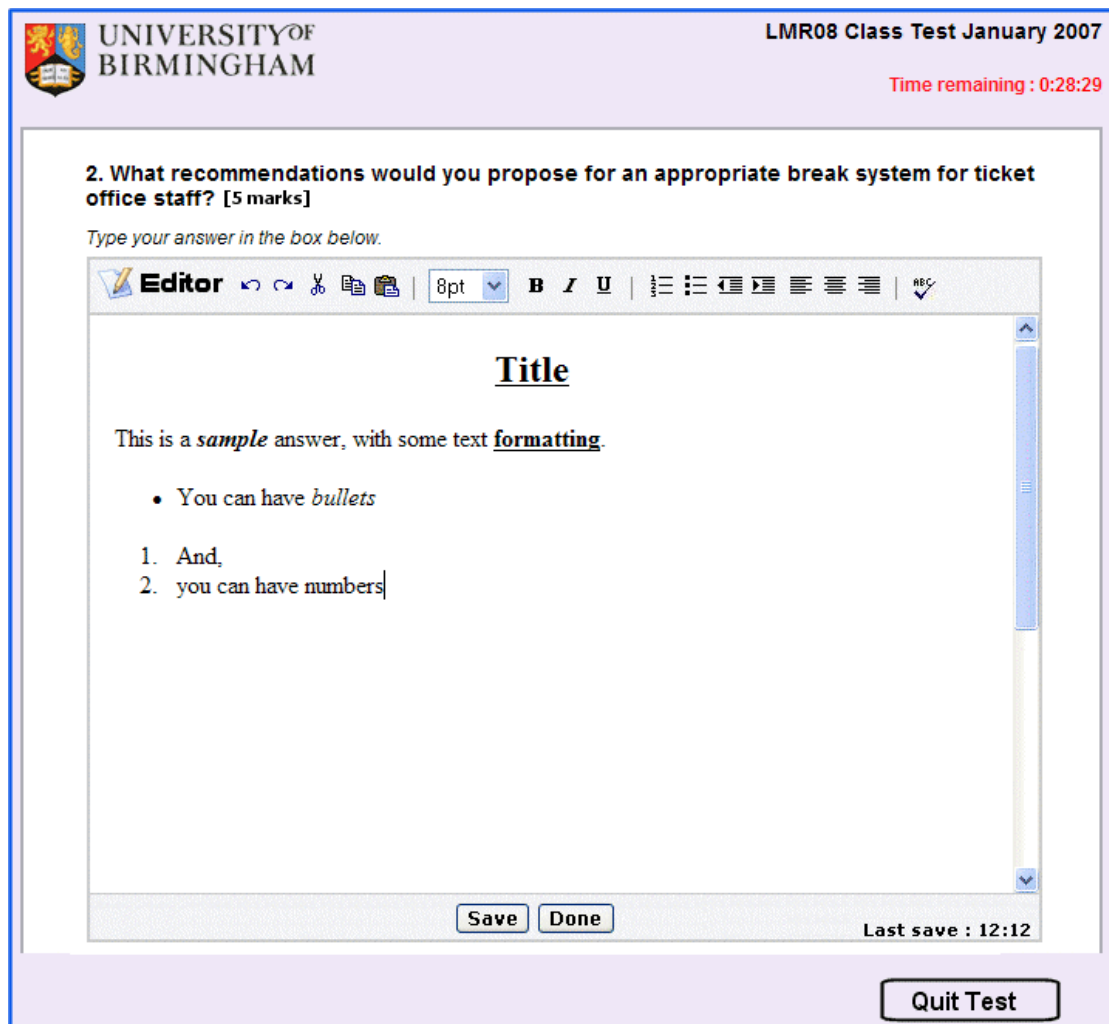
#### *The Student View*

The main examination screen for the candidates is the examination document previously authored by an administrator. Students answer each question by clicking on a link next to the respective question text (see Figure 1). This then provides an answer session for the question chosen.

For the individual answer session, the candidates are presented with a screen providing a simple and familiar word processor interface, supporting:

- Composing text;
- Copy/cut-and-paste;
- Font styles;
- Bulleting;
- Numbering;
- Text alignment.

A substantial amount of 'white space' is provided for the candidate's use, encouraging not only a discursive style of writing but also allowing the student to present his or her views appropriately. The screenshot shown in Figure 4 illustrates some of the basic formatting that can be used by students to structure their answers.



**Figure 4: Students use a secure web-based word processor interface to answer each extended answer or essay question**

The ExamOnline system is designed to go beyond the capabilities of standard Computer Assisted Assessment (CAA) systems in supporting essay and extended answer questions. The key features thus include:

- Copy/cut-and-paste, and simple formatting;
- Provision of an integrated spell checker (available at the administrator's discretion);
- An "autosave" functionality, taking a back up of student responses every 10 seconds or so.

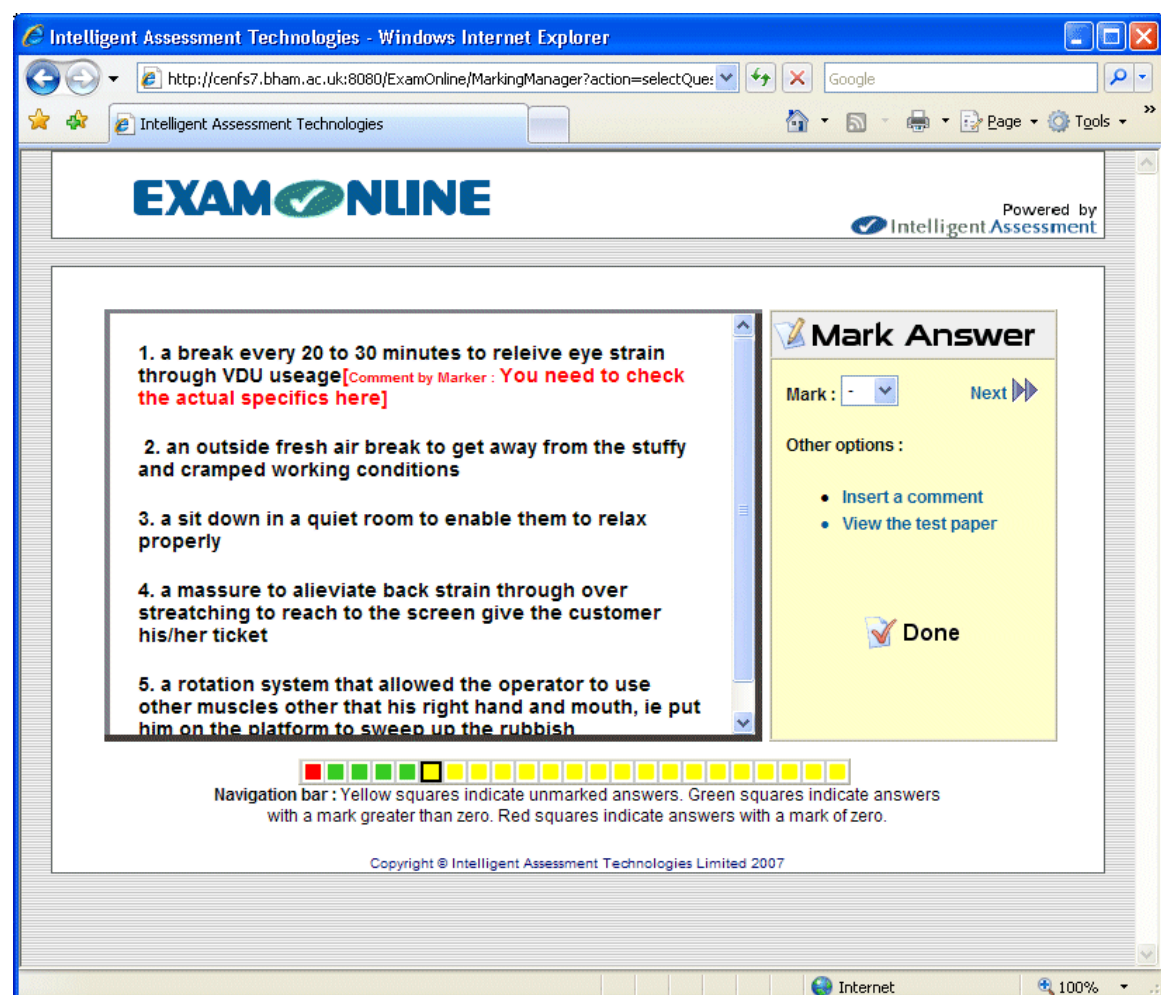
In addition, the interface has been designed to use asynchronous communication with the server. Thus it can cope with network and/or server outages, re-synchronising as and when the network comes back on-line.

### *On-Screen Marking of Examinations*

The designers of ExamOnline had in mind some key objectives when designing the marking interface:

- Make on-screen marking as simple and efficient as possible;
- Support existing marking practices (i.e., multiple markers);
- Support blind marking.

When markers log into ExamOnline, they are presented with a list of questions / papers to be marked. Clicking on a question brings them to the main marking interface that is shown in Figure 5.



**Figure 5: The markers' interface supports rapid blind marking, and also allows for the insertion of comments / feedback into student responses**

The interface has been designed to enable marking with the minimum number of mouse movements and/or keyboard input. For short or extended response

questions, using the keyboard provides extremely efficient processing of responses, with no mouse usage required at all. We estimate that marking is between two and three times quicker than would be the case for manual marking of scripts and, of course, here we are supporting blind marking.

In addition, the marking interface provides the ability to insert comments against each answer in the event that formative feedback to students is required, whether or not the output is also used in a summative manner. Such comments are shown as annotations (in red) against the student's response, as illustrated in Figure 5.

## **Security**

ExamOnline delivers examinations via a web browser. Accordingly, it is necessary to secure the browser so as to prevent student access to the internet, the local file system, email, etc. To ensure this level of security, students entering the ExamOnline system must first download and run a small Windows executable (itself delivered via the browser from the ExamOnline system). This executable:

- disables system keys (e.g., ctrl-alt-del, alt-tab, etc.);
- installs a 'keyboard hook' to trap browser 'hot-keys' which could otherwise be used to open new browser windows etc.;
- launches Internet Explorer in kiosk mode (that is, with no address bar, toolbars, or buttons visible or available) at the ExamOnline login page.

Once these actions have been carried out, candidates can only navigate and indeed exit the browser by using the interface provided by ExamOnline. Similar functionality is also available using commercially available secure browsers, such as Respondus LockDown Browser (Respondus, 2007).

Note also that, once logged in, students are unable to re-login without being provided with an additional invigilator password. Therefore, they cannot leave the invigilated environment and re-access the examination.

## **Robustness**

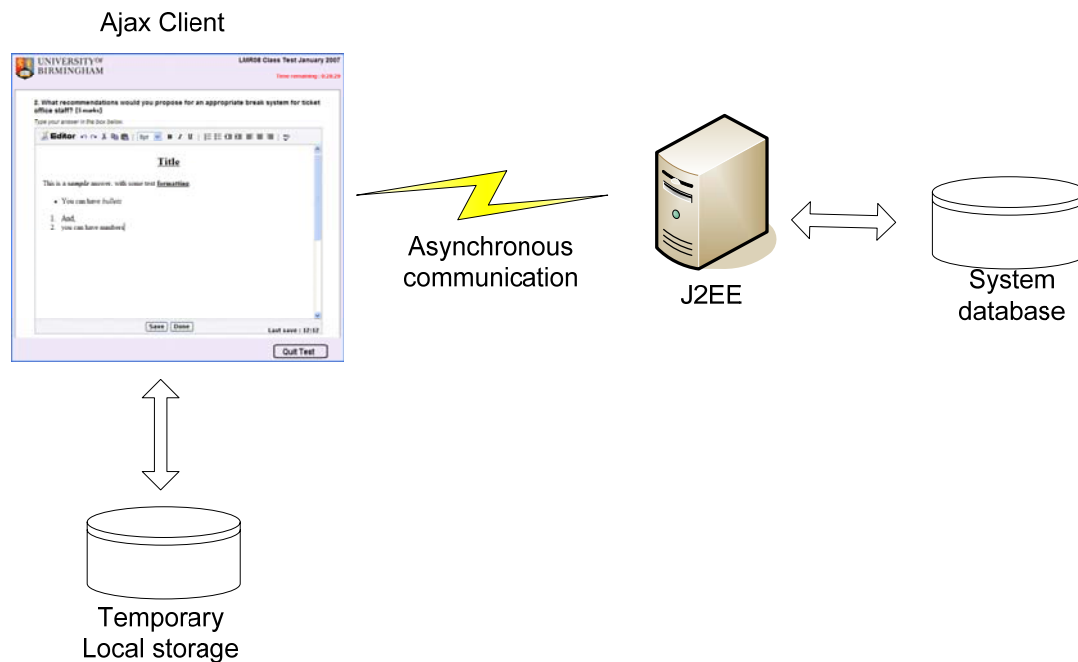
The student interface has been carefully designed to make the examination process as robust as possible. Specifically:

- After login, the entire test is downloaded to the student computer. No further communication with the server is required except to save student responses (and for spell checking, if enabled);
- There is an automatic 'autosave' functionality built in, which saves student responses to the server every 10 seconds or so;
- Student responses are saved to the server asynchronously. If the network / server is temporarily unavailable, a local copy of the



responses is kept and synchronisation is re-established with the server when the network / server becomes available again.

These and other features are implemented using an Ajax (Garret, 2005) client and a J2EE Java platform on the server, as shown in Figure 6.



**Figure 6: An Ajax client delivers the test to the student. This supports asynchronous communication with a J2EE server, providing robustness when network and / or server become temporarily unavailable**

## Results of Live Pilots

At the time of writing, ExamOnline has been used to deliver two live class tests with a largely identical student cohort. The system was modified and developed between the two tests, based on student and examiner feedback. Two academics were involved in setting the respective exam papers and in assessing the answers on-line.

### *Issues Relating to the Student Experience*

From observation of the examination sessions delivered to date on ExamOnline, the following points can be made:

- Students were initially cautious of taking summative examinations on-screen rather than on-paper. However, this effect was largely confined to their first encounter with the system. In a subsequent session, they 'just got on with it';

- The less computer literate students commonly expressed a desire to have a longer time period within which to complete the on-screen examination, compared to the paper version;
- It was noticeable that students readily took to re-visiting and re-editing responses to questions – much simpler and more effective in an electronic system than when using a paper-based system;
- A number of refinements to the interface were requested by the students after the first trial, and subsequently implemented. Further refinements are still outstanding (see further work);
- Accessibility was a problem for a student with very poor eye sight, who felt better able to view and answer questions on paper than on-screen. Some work on resolving this issue is planned for a later trial.

Students views were collected formally as part of a short survey conducted during the year-end examinations (see next section) and a selection of comments were received, as follows:

- I found the computer based test OK no problems.
- Ability to cut and paste similar text [useful].
- Adequate time should be given for those with limited typing skills and accuracy.
- I think the time and number of questions needs to be carefully considered.

Overall, the response from students has been favourable, noticeably more so after the second trial than after the first. This supports our view that, as these kinds of systems become more widespread and familiar, it is likely to be the paper-based examination process which will increasingly draw criticism from students.

### *Statistical Survey of Students' Views*

The course team conducted a brief survey of the students' experience of all their examinations in the academic years 2005/06 and 2006/07. This included paper based year end exams (lasting 2.25 hours) and the class tests (lasting 1.25 hours), two of which had been used for this trial. Three of the questions are relevant to the present paper:

1. Indicate the difficulty of each class test already completed (high, medium, low, do not know);
2. Indicate the suitability of the class test format in each case (high, medium, low, do not know);
3. How do you rate your computer skills? (high, medium, low, no answer)

The survey questions purposely covered both paper based and computer administered examinations to ensure that the team would be able to draw fair and comparative lessons. The responses from the 30 participants in the



survey were translated into numerical values (high = 3 and low = 1) and averaged, with the following results:

- Level of difficulty: The average for the two paper-based tests was 2.19 and that for the two computer-based ones was 2.22, both out of 3;
- Suitability of format: The average for the two paper-based tests was 2.17 and that for the two computer-based ones was 2.32, both out of 3.

The difficulty of both types of test was thus viewed as about right although the computer based format appeared to be slightly preferred over the paper-based one. The responses were then correlated with the students' own assessment of their level of computer literacy. For the first pilot, this indicated that students with a high level of computer literacy found the test more difficult than those stating a lower level of expertise, possibly indicating dissatisfaction with the MK1 user interface. The result for the second pilot, with a much improved interface, aligned expert users with a perception of lower difficulty.

Students were asked three further questions, but their answers were only analysed if they had taken part in one of the computer based tests:

- A. Would you be happy to be assessed in a class test in this way again?
- B. Would you be happy for all class tests to be run in this way?
- C. Would you be happy to sit a year-end exam in this way?

19 out of 23 respondents answered 'yes' to question (A), with 3 abstentions; 13 replied 'yes' and 7 'no' to question (B), also with 3 abstentions, while only 5 people would be happy to use the computer-based approach for an end-of-year exam (C), with 15 answering 'no' and 3 abstaining. Overall, the team feels that this outcome represents a positive result for the pilots.

It is worth noting again here that the MSc students who participated in these tests are all mature students, and few of them are what has been termed 'digital natives' (Prensky, 2001). We might reasonably expect therefore that the attitude of undergraduate students towards on-screen testing will be yet more positive (as has been found in other studies with younger participants (Sim, Horton, 2005)), and we hope to investigate this shortly.

#### *Issues Relating to the Examiners' Experience*

The two examiners involved in the trials had 12 and 8 years of experience respectively with assessing paper based class tests on this course. They both had concerns, initially, about the students' ability to cope with typing answers on-line. However, they had also seen a decline in students' ability to hand-write at the speed necessary to succeed. Their observations were as follows:

- The marking interface is user-friendly and requires only a minimal amount of training;

- Reading the answers is much easier than when having to decipher poor handwriting. Much ambivalence and subjective interpretation is removed from the assessment;
- Marking is fairer since the system allows marking question by question thus eliminating both positive and negative influences from poorly / excellently answered questions before and after. Both markers chose not to use the paper by paper assessment option;
- The progress bar gives a very positive indication to the marker as to how much has been achieved and how much is still to do...

Feedback to the designers of ExamOnline resulted in a number of modifications to the interfaces between the first and second live trial and also in many of the suggestions for further development, discussed later on in this paper. Overall, they declared themselves very satisfied with their own experience of using the system.

### **Future Work Planned**

There are a number of areas where ExamOnline is currently being improved on the basis of the two live pilots, and in preparation for further roll-out :

- Inclusion of differentiated mark schemes for individual questions, which will be integrated into the marking interface;
- “Offline marking”, to support “as and when” marking on personal computers and laptops, with later synchronisation with the main system;
- Integration with back end systems for outputting results;
- Integration with a free-text computerised marking system to provide automatic marking of short answer questions (Intelligent Assessment Technologies, 2007);
- Support for drawing diagrams when answering questions, potentially on-screen (Thomas, 2004), but with an option for hand-drawing and paper based submission of calculation-steps;
- Enhanced accessibility for sight impaired students;
- The ability to build up (and insert from) a list of standard comments as marking of a question progresses;
- Addition of simple QA measures into the marking process (e.g., item statistics);
- Support for double marking of responses.

### **Conclusions**

A new e-assessment system, ExamOnline, has been specifically developed to deliver summative, essay style examinations in an HE/FE setting. The system has been designed to support existing examination processes, to provide a

better and more relevant examination experience for an increasingly digital cohort, and to support an efficient blind marking process. Initial pilots have confirmed that the system provides an effective and efficient means of deploying traditional essay style examinations on-screen and that it improves in many ways upon the existing paper-based process. The system will undergo further development and roll-out in the coming months, based on the feedback received during continuing live pilots with students on a Masters programme. Further pilots with undergraduate students are planned for the coming months.

## References

Conole (2005), G. Evaluation Of The Scottish Pass-It Assessment Project, *9th International Conference on Computer Aided Assessment, Loughborough University, Loughborough, 2005*.

Garret, J. (2005) Ajax: A New Approach to Web Applications. Adaptive Path. Accessed on 11<sup>th</sup> May 2007.

<http://www.adaptivepath.com/publications/essays/archives/000385.php>

Intelligent Assessment Technologies, 2007.

<http://www.intelligentassessment.com>

Prensky, M. (2001). Digital natives, digital immigrants. *On the Horizon*, 9(5), 1–2.

<http://www.marcprensky.com/writing/Prensky%20-%20Digital%20Natives,%20Digital%20Immigrants%20-%20Part1.pdf>

Respondus 2007. Respondus LockDown Browser.

<http://www.respondus.com/products/lockdown.shtml>

Sim, G. & Horton, M. (2005). Performance and Attitude of Children in Computer Based Versus Paper Based Testing. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2005* (pp. 3610-3614). Chesapeake, VA: AACE.

Thomas, P. (2004). Drawing diagrams in an online examination. *8th International Conference on Computer Aided Assessment, Loughborough University, Loughborough, 2004*.

# **BLENDED DELIVERY MEETS E- ASSESSMENT**

**Eric Shepherd**



# Blended Delivery Meets e-Assessment

Eric Shepherd  
CEO, Questionmark

## Abstract

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 to work collaboratively.

Over the past year, Questionmark has introduced dozens of new e-assessment capabilities that have made true “Blended Delivery” a reality. This session will explain and demonstrate some of the new technologies that can help academic and assessment professionals author, deliver, monitor, and report on an increasing number of assessments easily and securely including:

- **Printing and Scanning:** Develop your test online, deliver on paper; scan in results for scoring and reporting.
- **Disconnected Delivery:** Extending the benefits of “online” delivery to “offline” students. Find out how Questionmark to Go enables e-Assessment delivery and results reporting for students who are disconnected from the Internet.

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 use assessments to measure knowledge, skills and attitudes.





# **DOMAIN-SPECIFIC FORMATIVE FEEDBACK THROUGH DOMAIN-INDEPENDENT DIAGRAM MATCHING**

**Christos Tselonis  
John Sargeant**



# Domain-Specific Formative Feedback through Domain-Independent Diagram Matching

Christos Tselonis, John Sargeant  
School of Computer Science  
University of Manchester  
{tselonic,johns}@cs.man.ac.uk

## Abstract

As part of our Human-Computer Collaborative (HCC) approach to assessment, we seek representations of answers and marking judgements which can be applied to a wide variety of situations. In this paper we introduce such a representation, which we call a *gree*<sup>1</sup>, and discuss an initial practical application of grees for formative feedback. An experiment was carried out in which students were asked to construct an answer while receiving interactive feedback and then complete a short survey. The results show that it is possible to give effective domain-specific formative feedback based on a domain-independent internal representation or “metaformat”.

This work builds on results we have previously presented on domain-independent diagram matching based on heuristic matching of graphs. Grees provide much greater flexibility, with a wide variety of potential applications. We discuss some problems which need to be overcome before we can realise their full potential.

## Introduction

Fully automated marking for constructed answers such as diagrams and text is a very difficult task. Although there are implementations attempting to generalise the marking process [2], most efforts focus on single knowledge domains or depend on particular semantics [1, 8, 9, 11], lacking reusability and extendibility.

We have proposed the human-computer collaborative (HCC) approach as a solution [7], according to which marking is a dynamic process where the computer deals with repetitive tasks while the human makes the important judgements. We have shown that such an approach can significantly reduce the effort and the time taken for a human to mark a large number of answers.

---

<sup>1</sup> As part of the commercialisation of the ABC software by Assessment21 Ltd., the use of grees in assessment is the subject of a patent application.

In parallel, we now attempt to enhance the student experience, by extending the system to dynamically generate real time feedback, based on matching the student's answer against a model answer.

## The story so far

In [10] we discussed a way of matching constructed answers, and in particular diagrams, based on heuristics. The method involves the conversion of answers into enriched graphs, whose components (which we call “boxes”, and “connectors”) retain some of the attributes existing in the original diagrams, such as types and labels (text strings associated with the boxes and connectors). Feeding a model graph and any number of student graphs into the matching mechanism, along with a set of metric / weighting pairs determining the matching process, results in a number of local scores associated with the graphs' nodes, which eventually are combined to produce a score of similarity between the graphs.

The similarity scores not only resemble -in most cases<sup>2</sup>- the marks previously awarded by a human marker, but most importantly could provide the means to improve consistency and minimise marking time; sorting the answers by similarity to the model answer or viewing these similarities highlighted in colour certainly helps in this respect.

In the next section we explain grees and gree matching. Then we describe an encouraging experiment in using grees for formative feedback. Finally we draw conclusions and discuss a number of further issues.

## Grees and the matching mechanism

The revised matching mechanism, although based on the one introduced in [10]<sup>3</sup> includes significant enhancements; it is now extended to adopt a *modular scoring strategy*, according to which *parts* of the model answer are separately matched to parts of the candidate answer. This way, marking schemes can be accurately defined and marks awarded for the parts of the answer that really matter, although they can be dynamically amended later on if necessary. Different parts of the model answer may be worth a different portion of the total marks available, and can also be weighted differently, according to their components' relative importance. Equally importantly, multiple alternative acceptable parts deserving the same portion of marks can be set for a single constructed answer.

To enable this modular approach, *grees*, **dynamically extendable AND/OR trees whose leaf nodes are overlapping graph fragments**, were invented. They effectively represent the model answer parts along with any marking

---

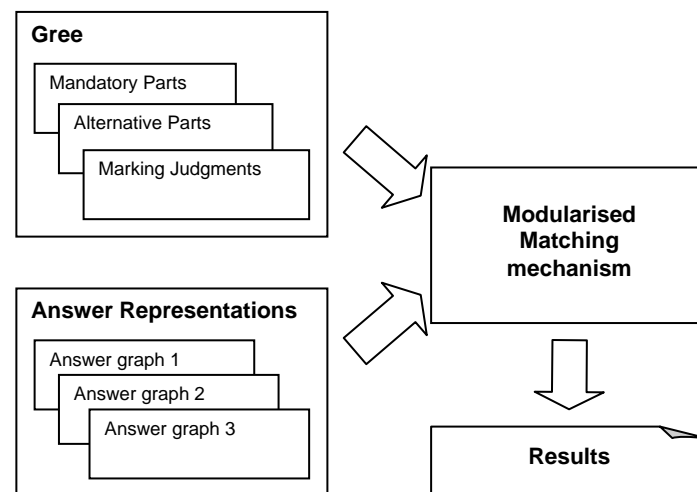
<sup>2</sup> Cases where student drawings abided by some basic rules, e.g. an answer should be a single connected graph, box labels should be placed *in* the box, not above it etc.

<sup>3</sup> A number of details, explained in [10] are omitted from the description here, so readers desiring a complete account of the matching process will need to consult that paper

judgements and other information needed in a systematic manner, allowing for reusability, extendibility and modularity. Although some aspects of grees, such as the use of AND/OR trees to represent marking schemes, have been proposed before, the combination is, to the best of our knowledge, novel.

Being a generic metaformat, grees do not depend in any way on the knowledge domain of the question; as long as an answer can be converted to a graph consisting of boxes and connectors, any answer type may be modelled by a gree, including diagrams, mathematical expressions, software programs and even short, factual text fragments. No domain-specific information is contained in a gree, or used by the matching mechanism.

The matching process takes place between a model answer stored in the gree metaformat and a set of candidate answers converted to graphs as shown in Figure 1.



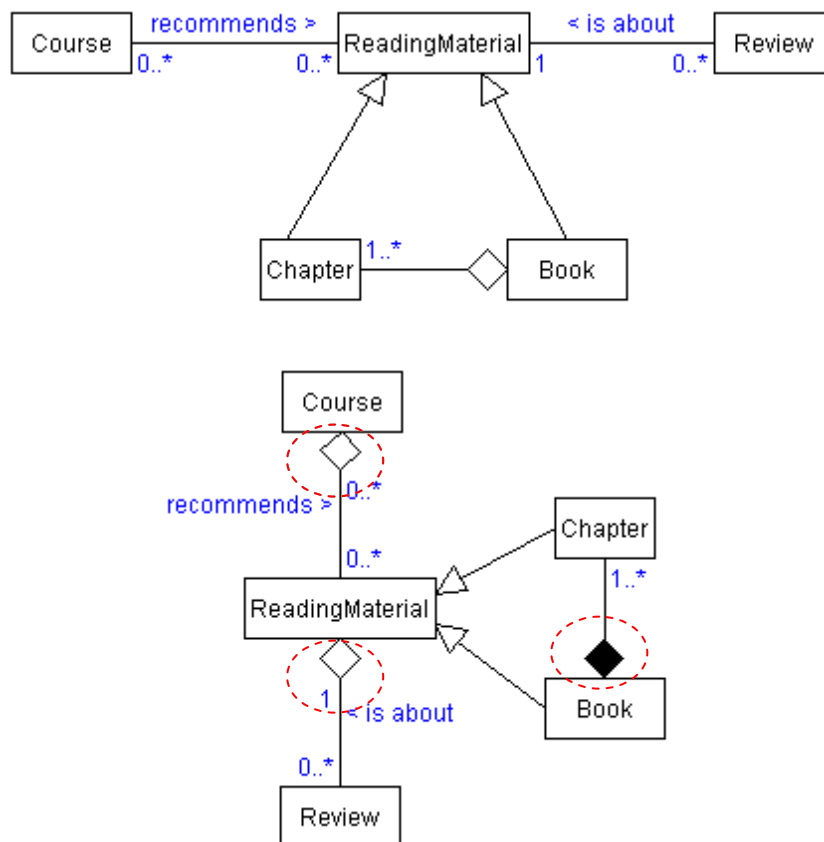
**Figure 1: the matching process.**

## Grees in detail

Our example is based on the question in listing 1, set by the second author for a software engineering examination in January 2006, which requires students to draw a UML class diagram. Although not trivial, the question tightly constrains what a correct answer must look like. Figure 2 shows two possible fully correct answers. They include several different components (circled), but also some spatial differences, which are ignored by the matching mechanism.

You are designing an online book information system for the Resource Centre. This will allow students to find out about which books, or chapters of books are recommended for each course, and also to read or write reviews of books or chapters for the benefit of other students. The software will also attempt to provide a summary of each Chapter. You have identified the domain classes *Book*, *Chapter*, *Course*, and *Review*, and there will be corresponding design classes. Since Chapters as well as complete Books may be recommended or reviewed, you have added an additional design class *ReadingMaterial* to capture the common properties of the two. Draw a skeleton design class diagram to show the exact relationships between these five classes (but not their attributes or operations).

**Listing 1: The examination question.**



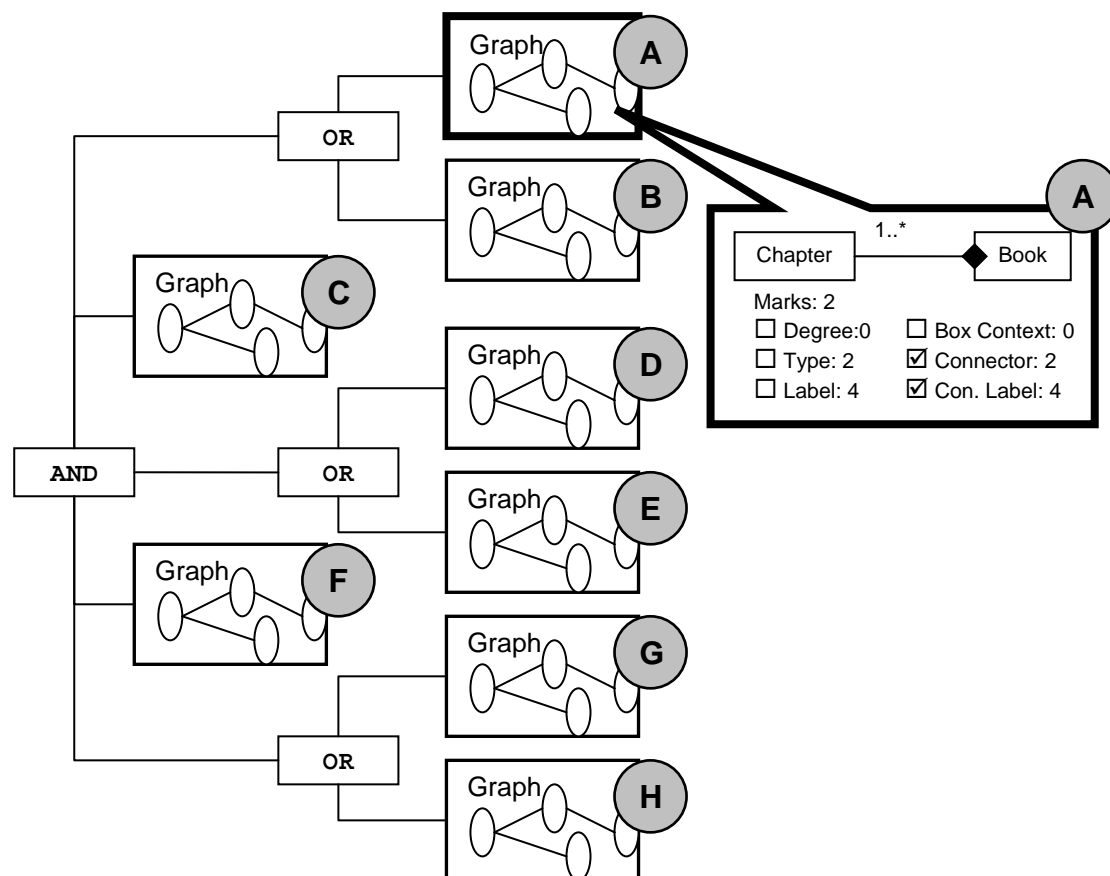
**Figure 2: Two of the possible fully correct answers.**

Figure 3 represents a tree which describes 8 fully correct answers to the question. Its leaf nodes, labelled A to H, contain possible parts of the correct answer. Each part answer comprises a portion of an acceptable class diagram, in combination with a number of parameters describing how the answer part should be matched. In particular, the parameters include the number of marks allocated for the answer part, the generic metrics considered during the matching process for the various components, and features comprising the answer part, weightings specifying by how much the metrics should count towards the final score and flags determining whether these

metrics should contribute to both the matching and the final score, or just the former, in order that the algorithm matches corresponding parts correctly.

For example, consider node A in Figure 3. In terms of the question, this specifies one of two possible correct ways of representing the relationship between the Chapter and Book classes (node B being the other). In particular the black diamond, representing a strong association, is important.<sup>4</sup>

The algorithm must first ensure it is matching the correct part of the graph, in this case the boxes labelled Chapter and Book and the connection between them. So for instance the box label metric has the maximum weight of 4. Having made the correct match, for scoring purposes we only care about the type of connector and its label (as indicated by the ticks in the "check boxes"). Note that between the previous paragraph and this one we have moved from domain-dependent concepts to a domain-independent algorithm.



**Figure 3: A gree representing a set of correct answers.**

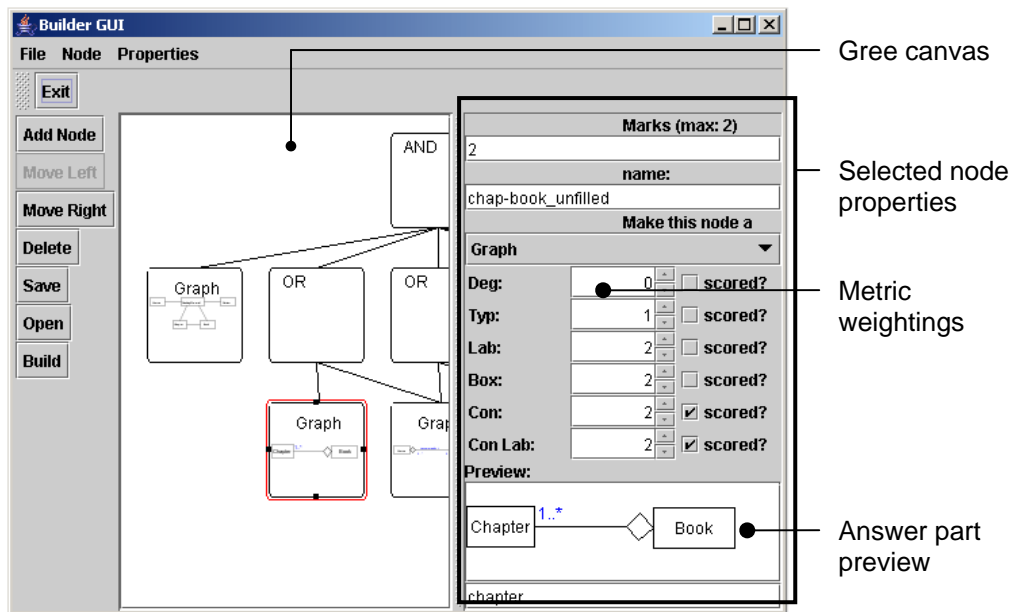
<sup>4</sup> Although the first author's original marking scheme only allowed for the option represented by node B, an example of the HCC principle that marking schemes usually need to be extended dynamically on the basis of student answers.

The leaf nodes are connected in a tree structure by AND and OR nodes. For a submitted answer to be awarded full marks, it must contain all the sub-parts specified by the subtree below an AND node. For a group of leaf nodes placed directly under an OR node, the content of only one of them need form part of the submitted answer. Clearly in this example, a submitted answer worth full marks must contain (A OR B) AND C AND (D OR E) AND F AND (G OR H). This gree fully describes  $2^3 = 8$  alternative and fully correct answers.

In order to match a submitted answer, such as either of those shown in Figure 2, against a gree, the matching mechanism starts by considering the gree's root node and continues traversing the nodes down the tree. Once a leaf node is encountered, i.e. a node that contains a graph fragment, a score for that node compared to the submitted answer is calculated using heuristic methods as explained in [10]. For all nodes descending from an OR node, the one producing the highest score is considered to be the closest match. This score is thereafter the one associated with that OR node. For all nodes depending directly from an AND node, the scores are added. Once the matching process has completed, the score given by the root node is the mark awarded to the submitted answer.

Theoretically, grees can be re-adjusted on the fly during the marking process in the light of previously unconsidered alternative correct parts of submitted answers. This could involve adding more nodes, reconnecting existing ones differently, splitting the marks differently, or changing values for the metrics. The system will then automatically recalculate the scores for the already marked answers and notify the human marker for the submissions whose scores have changed. The gain can be significant over traditional paper based marking, where changes to the marking scheme part way through a large number of submissions require reviewing all answers marked so far. Although the gree specification supports it, a tool which would allow end-users (i.e. markers) to edit grees has not yet been developed, since the user interface issues are significant. However, an experimental editor application exists (Figure 4), which may form the basis of a marking tool in the future.

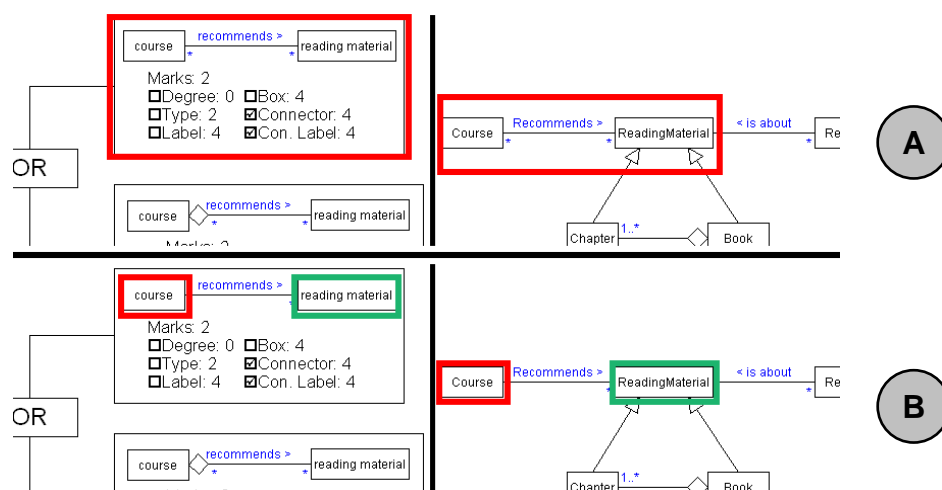




**Figure 4: A basic gree editor.**

According to the HCCA paradigm, the human marker is responsible for affirming or amending the automated marking results. To ease this process, a marking tool based on grees could support the visual features discussed previously [10]; a part of the submitted answer can be highlighted with the same colour as a gree's node, to indicate a match (Figure 5A). Alternatively, the contents of the gree's nodes, which are live graph objects, may be coloured according to matched parts of the submitted answer (Figure 5B).

Additionally, sorting the submitted answers by mark, status (marked / unmarked), completeness (number of gree nodes matched) etc is a straightforward extension.



**Figure 5: Communicating the matches visually.**

Another important feature of the gree is the ability to reconstruct programmatically the full set of possible correct answers. In addition they provide the ability to determine which of these correct answers is the closest to another answer. This is important when using grees to provide students with instant feedback, for example during formative or self assessment

### Use of grees in formative assessment

As a first practical application, a standalone tool intended for formative and self assessment [2, 3, 5, 6], taking advantage of the gree matching mechanism, was developed<sup>5</sup>. It displays a question to a student, allowing them to draw the answer, while providing automated feedback. The tool translates the results of matching the student's current drawing against a gree into meaningful feedback strings. The strings, which may vary from general hints to very specific information such as suggested content and component locations, are displayed on the drawing canvas via popups. Listing 2 displays a number of example feedback strings.

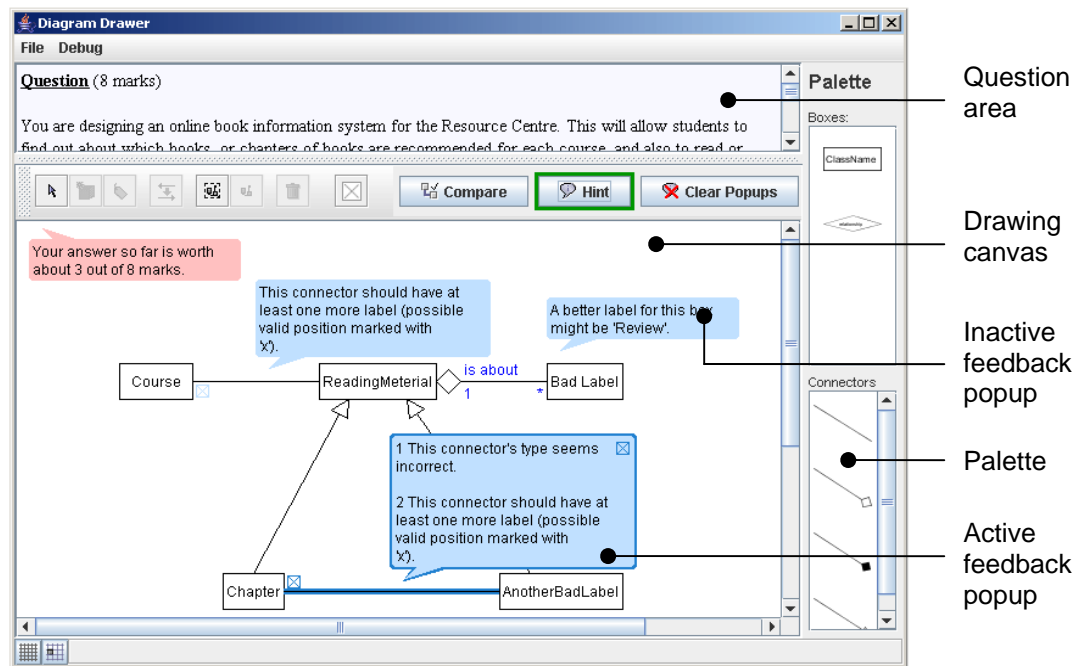
- A better label for this box might be '*Course*'.
- This box's type seems incorrect (Should be '*Class*').
- There are 2 boxes too many connected to this one.
- There should be 2 more boxes connected to this one.
- This connector's type seems incorrect.
- This connector's direction seems incorrect.
- This connector should have at least one more label (possible valid position marked with 'x').
- This connector has at least one label too many.
- One or more of this connector's labels are misplaced (possible valid position marked with 'x').

#### Listing 2: Example feedback strings.

Figure 6 displays the feedback tool in action. Hovering over a popup “activates” it, highlighting the popup as well as the component it refers to. Clicking on it causes it to be dismissed. An extra button to clear all popups at once is also available.

---

<sup>5</sup> Based on a diagram drawing tool initially developed by Stuart Anderson.

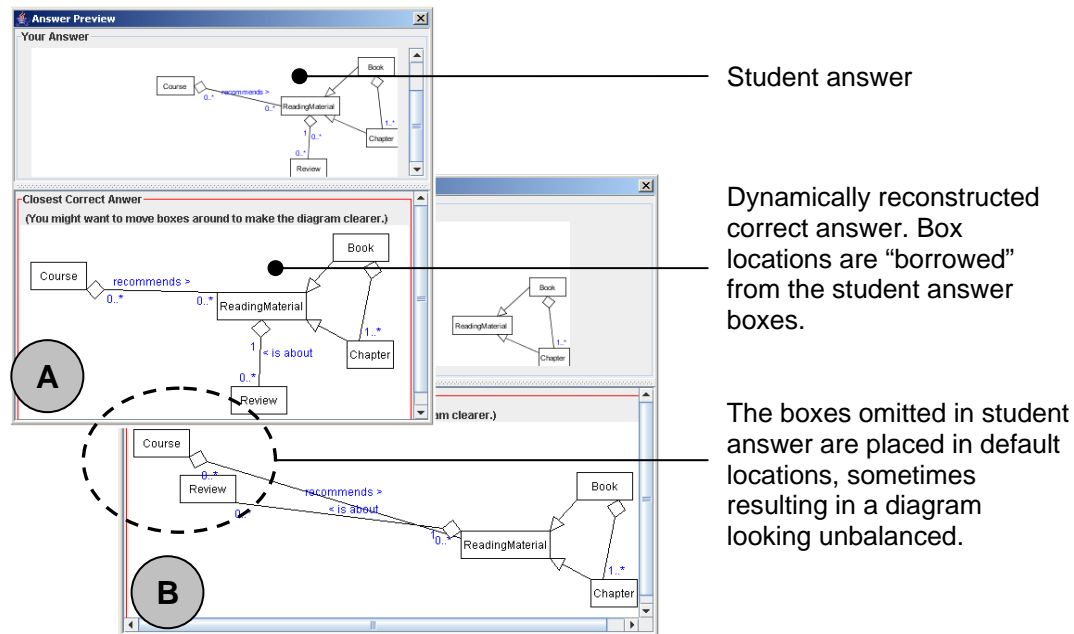


**Figure 6: The feedback tool.**

The tool also incorporates a 'Compare' button which when pressed, will query the system for the fully correct answer which is closest to the current drawing. It will then display both side by side in a new window. Figure 7A displays the closest correct dynamically reconstructed answer to a student answer shown at the top, while Figure 7B displays the same diagrams, with two of the student answer's boxes omitted. In this case, the closest answer looks somewhat unbalanced. Students, however, were able to drag the boxes around to make the diagram clearer.

## The experiment

Second year Computer Science students attending the '*COMP2341: Software Engineering I*' module were asked to take part in this experiment, evaluating gree matching and feedback generation. The feedback tool was deployed as a Java applet, capable of running over the Internet on any Java-enabled browser, so students could run it in their own time, completely anonymously. They were also given the option of a supervised session following an exam revision class, but none made use of this opportunity, possibly because it was four whole days before the exam. Therefore all students who participated did so with no help or supervision.



**Figure 7: Dynamically reconstructed correct answers.**

First, the trial featured a short tutorial session during which students had the opportunity to familiarise themselves with the tool and the feedback mechanism, by following a series of step-by-step instructions in order to construct a trivial diagram. During the tutorial, the students were guided to intentionally make errors so the feedback features, triggered automatically upon their actions, were emphasised.

They were then presented with the main question, (Listing 1) from the previous year’s examination, asking them to draw a UML class diagram like those shown in Figure 2. Both the feedback and the ‘Compare’ buttons could be used at any time, any number of times. However, all such interaction was being recorded and when viewing the closest fully correct answer for comparison, editing the answer was disabled.

Once a student elected to commit to their final answer, they were presented with a short, optional, survey, assessing the tool’s usefulness. The survey responses along with the diagram answer and the statistical data were finally submitted back to the server.

## The results

A total of 42 submissions were received, two of which contained no usable data. Although there was no definitive way to determine whether all submissions were submitted by different users because of complete anonymity, it is likely that all or most were, judging from the differences between the answers and the submission timestamps.

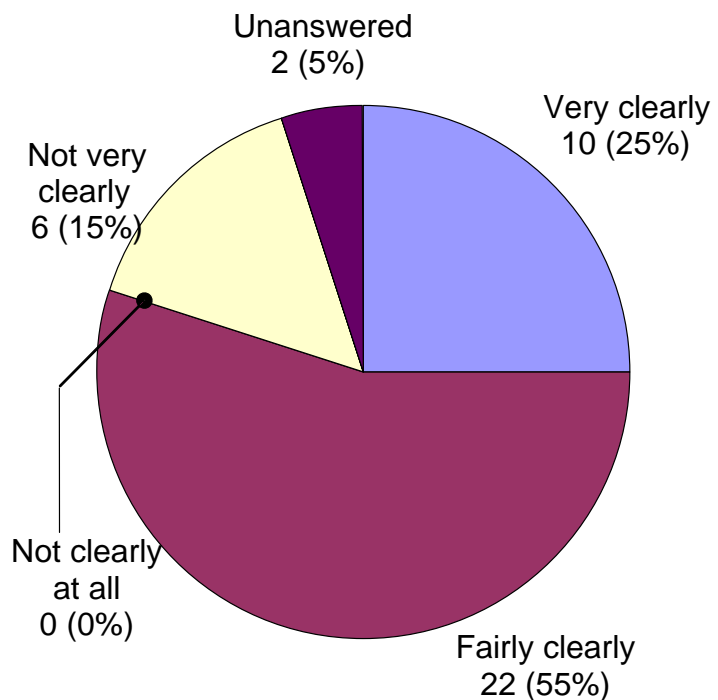
The first question of the survey was “*How many times did you use the hint mechanism?*”. The students could enter an integer in a spin edit control, or leave the default 0. The number of times the feedback mechanism was *actually* used was recorded and ranged from 0 to 60 per submission. The difference between the survey responses (estimates) and the actual number of times was great, both overall and on a per student case; generally, students tended to underestimate this number by about a factor of 2. Table 1 summarises the estimated and actual ranges. For instance 25 students believed they had used the feedback mechanism no more than 4 times, but only 9 had actually done so.

Times Used	Submissions	
	Actual	Estimate
0 - 4	9	25
5 - 9	12	6
10 - 14	3	3
15 - 19	2	4
20 - 24	5	0
25 - 29	3	0
30+	6	2

**Table 1: Actual and estimated number of times the feedback mechanism was used per submission.**

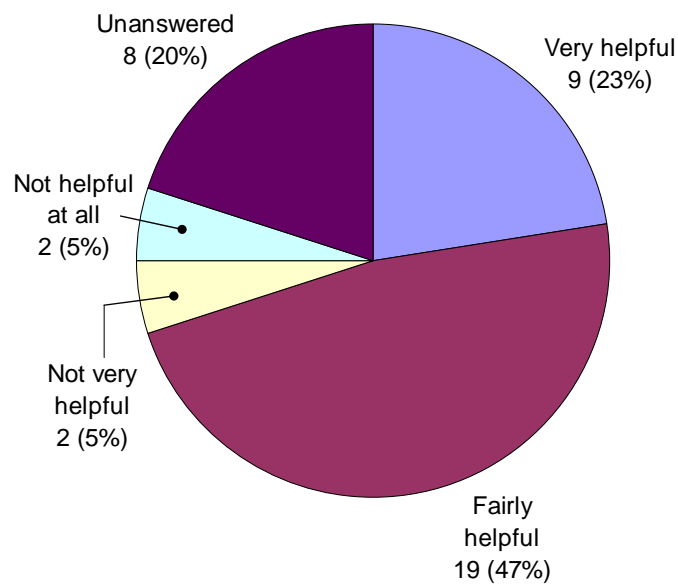
The second survey question was “*How clearly was the feedback information presented?*”. The students could select one of four options, shown in Figure 8. According to 32 submissions (80%), the feedback was presented fairly, or very clearly.

The third question was “*How helpful was the feedback received?*”. Similarly to the second question, the students had a number of options to choose from (Figure 9). According to 28 of the submissions (70%), the feedback was fairly, or very helpful, while a 20% did not answer this question.



**Figure 8: Answers to question “How clearly was the feedback information presented?”.**

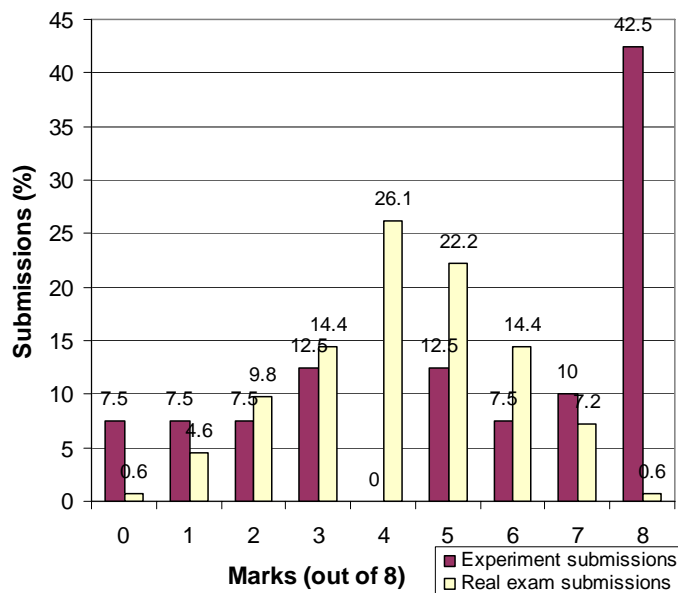
The last question was “*What would you suggest to make the feedback mechanism better?*”. A text area allowed the students to enter text of any length. Listing 3 displays the responses (14, since there was no response in the rest of the submissions) to this question. It is worth noting that the 5 students who responded purely positively in this question (cases 6, 7, 12, 13 14), were awarded high marks. The comments including constructive feedback touched mostly issues with the mechanism, that were known in advance. For instance, label matching (4, 8) and popup positioning (5, 10, 11) were not optimal. Additionally, some of the suggested defects were intended that way. For example, the message in the first comment is displayed whenever the current answer is almost identical to one of the specimen solutions stored in the gree, hence there is no useful feedback to be provided, although the message, could be clearer.



**Figure 9: Answers to question “How helpful was the feedback received?”.**

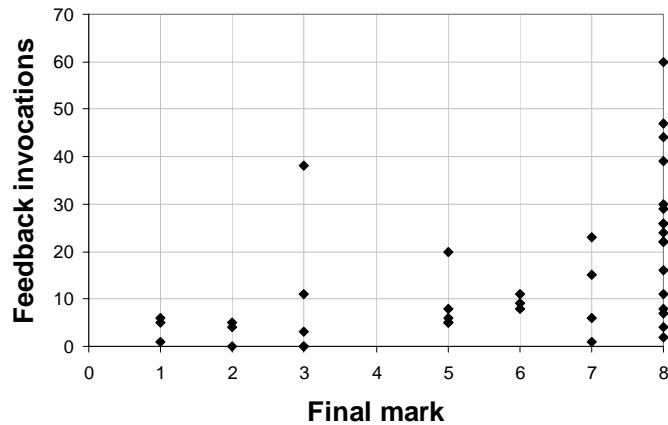
Comment 9 is interesting because the tool was trying to direct the student towards the right answer - labels such as 1..\* are never placed in the middle of a connector in UML - but the student was refusing to be helped!

The maximum marks available for the question were 8. 21 of the submissions (52.5%) were given an estimated mark, based on gree matching, between 7 and 8 marks. Figure 10 displays the marks the final submissions were awarded, compared against the marks awarded by a human marker for the real examination, a year earlier, when only one out of 153 students received full marks. Obviously, when using the feedback mechanism, the marks tend to be higher, while for the cases where the marks



**Figure 10: Final mark allocations.**

were low, the feedback system was barely used and the question was probably abandoned half way through. Figure 11 shows that generally, the fewer the times the feedback mechanism was invoked, the lower the final mark. However, the lower right corner of this plot shows a number of high marks with relatively few hints, suggesting that the experiment prompted a number of students to do extra revision before using the tool.



**Figure 11: Marks – Feedback invocations correlation.**

- 1. Really quite good. Bug... 'No Feedback could be generated this time.' repeats. (8 marks, 4 hints)
- ↓ 2. Have it make sense. (3 marks, 3 hints)
- 3. Better descriptions (2 marks, 4 hints)
- 4. was very exact about names, it didnt recognise Reading material it wanted it without a space and Recommends > was told it shud be called recommends > (8 marks, 60 hints)
- 5. sometimes they overlap which can be abit confusing/annoying. Maybe some kind of list of hints? like view next hint or something. Don't know if was intended but hints can just be used repeatedly to find the answer, but maybe that was the point? Also, I have no idea how many times I used hint.. It was lots. Very helpful anyway (7 marks, 15 hints)
- ↑ 6. Don't really know, its good at the moment and helped loads cheers :) (8 marks, 7 hints)
- ↑ 7. Nothing seems fine as it is (7 marks, 1 hint)
- 8. More intuitive suggestions, i.e. maybe more correct answers for it to choose from? The problem I had was that it would suggest that some of my correct aspects were incorrect and confuse me by telling me it was incorrect. (6 marks, 9 hints)
- 9. Include more flexibility for labeling syntax. Such as allowing \*..1 to be placed in the middle of the connection. (5 marks, 5 hints)
- 10. Dont overlay the feedback boxes (8 marks, 39 hints)
- 11. Pop ups are a bit annoying, maybe have a feedback area and when feedback is clicked on area that needs changing is highlighted. (8 marks, 2 hints)
- ↑ 12. No need to improve (8 marks, 22 hints)
- ↑ 13. It's fine as it is (7 marks, 23 hints)

↑ 14. More questions to tackle with detailed feedback (5 marks, 5 hints)

Key: ↑ Purely positive comments  
→ Constructive feedback  
↓ Purely negative comments

### Listing 3: Survey responses to the question “What would you suggest to make the feedback mechanism better?”.

A similar question was set in the real exam in January 2007, and many more received high marks (7 or 8) compared to the previous year. This cannot be primarily due to the feedback tool, as less than a third of the students participated, in the experiment, but it suggests that the feedback tool may have had a significant positive effect for some students. Figure 12 compares the marks awarded during the two examination runs.

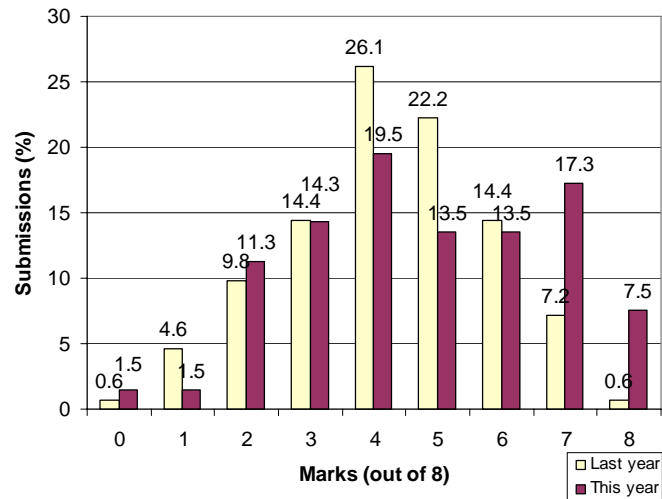


Figure 12: Comparison between the two examination runs.

To check that the estimated marks given by the feedback tool were reasonable, the matching mechanism was tested on a random sample of 48 out of 153 answers received for last year's examination. The automated marks were compared against the marks the human awarded. The results, shown in Figure 13 also indicate that the gree method has the potential to work effectively as a human marker's guide. Discrepancies are largely due to problems with label matching. For instance the

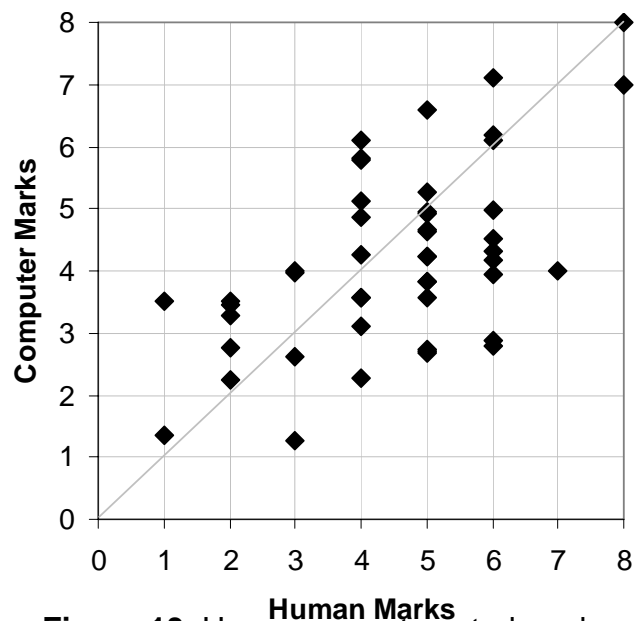


Figure 13: Human vs. automated marks.



labels “ReadingMaterial” and “Reading Material” are treated as different. Use of edit distancing and other techniques used in analysing text answers is required to address this issue.

## **Conclusions and further work**

Grees, a special metaformat to represent structured answers including alternative parts and marking judgment information, independently of the knowledge domain, were introduced. They can be dynamically extended and in combination with a modularized matching mechanism, comparing and matching a submitted answer against a model answer is possible. The results can be expressed both visually and in terms of estimated marks.

The system was extended to include another mechanism that dynamically translates the matches into feedback combining explicit strings and visual information. A client application employing the whole system was deployed and 2<sup>nd</sup> year CS students were asked to answer a previous year examination question by drawing a UML diagram using it.

Although the trial group can be considered to be demanding given their exposure to computer systems, the experiment results proved to be clearly encouraging. The feedback mechanism was used several times per student and their final marks, compared to the ones from the examination the previous year, were significantly improved; in general, the more the feedback queries, the higher the final mark. According to the majority of the students, the feedback was at least “fairly helpful” and was presented at least “fairly clearly”. Some known problems, such as the weak label matching and the relative positions among the feedback popups, were pinpointed.

Since the reviewed draft of this paper, a second trial has been conducted, with first year AI students drawing Markov Chain diagrams. Although the type of diagram was quite different, the student feedback, both qualitative and quantitative, was very positive, and remarkably similar to that described above. This strongly reinforces the claim that a domain-independent representation can be used to give effective domain-specific formative feedback.

Future plans include testing the feedback system in other knowledge domains and even different types of constructed answers, such as mathematical expressions. Providing a user interface which allows markers to build and extend grees in an intuitive way remains an interesting challenge.

## References

1. F. Batmaz and C.J. Hinde. A diagram drawing tool for semi-automatic assessment of conceptual database diagrams. In *10<sup>th</sup> International Computer Assisted Assessment Conference*, 2006.
2. J. Burtner, R. Rogge and L. Sumner. Formative assessment of a computer-aided analysis center: plan development and preliminary results. In *Frontiers in Education*, 2004.
3. P. Bocij and A. Greasley. Can computer-based testing achieve quality and efficiency in assessment? In *International Journal of Educational Technology*, 1999, vol 1.
4. C.A. Higgins and B. Bligh. Formative computer based assessment in diagram based domains. In *Innovation and Technology in Computer Science Education (ITiCSE) 2006*, 2006.
5. David J. Nicol and Debra Macfarlane-Dick. Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. In *Studies in Higher Education*, Vol 31, No. 2, April 2006.
6. DR Sadler. Formative assessment: revisiting the territory. In *Assessment in education* 5.1, 1998.
7. John Sargeant, Mary McGee Wood and Stuart M. Anderson. A human-computer collaborative approach to the marking of free text answers. In *8<sup>th</sup> International Computer Assisted Assessment Conference*, 2004.
8. Neil Smith., Pete Thomas and Kevin Waugh. Interpreting imprecise diagrams. In *Diagrams 2004 Conference*, 2004.
9. Pete Thomas. Drawing diagrams in an online examination. In *8<sup>th</sup> International Computer Assisted Assessment Conference*, 2004.
10. Christos Tselonis, John Sargeant and Mary McGee Wood. Diagram matching for human-computer collaborative assessment. In *9<sup>th</sup> International Computer Assisted Assessment Conference*, 2005.
11. Athanasios Tsintsifas. *A framework for the Computer Based Assessment of Diagram Based Coursework*. PhD thesis, School of Computer Science and Information Technology, University of Nottingham, 2002.

# **OPEN MENTOR: SUPPORTING TUTORS WITH THEIR FEEDBACK TO STUDENTS**

**Denise Whitelock and Stuart Watt**



# Open Mentor: Supporting Tutors with their Feedback to Students

Denise Whitelock,  
Open University,  
Walton Hall, Milton Keynes  
[D.M.Whitelock@open.ac.uk](mailto:D.M.Whitelock@open.ac.uk)

Stuart Watt  
The School of Computing  
The Robert Gordon University  
Aberdeen  
[s.n.k.watt@rgu.ac.uk](mailto:s.n.k.watt@rgu.ac.uk)

## Abstract

Assessment is one of the major challenges for higher education today. This is partly because it traditionally squares the desire for improved constructivist learning against the demand for institutional reliability and accountability. The call for a pedagogically-driven model for e-Assessment was acknowledged as part of a vision for teaching and learning in 2014 (Whitelock and Brasher 2006). Experts believe that such a model will allow students in Higher Education to take more control of their learning and hence become more reflective. These are indeed laudable aims but how can they be implemented in practice?

One of the problems with tutor feedback to students is that a balanced combination of socio-emotive and cognitive support is required from the teaching staff, and the feedback needs to be relevant to the assigned grade. Is it possible to capitalise on technology to build training systems for tutors in Higher Education, that will support them with their feedback to students, and which will encourage their students to become more reflective learners?

## Introduction

One of the challenges of today's education is that students are expecting better feedback, more frequently, and more quickly. Unfortunately, in today's educational climate, the resource pressures are higher, and feedback is often produced under greater time pressure, and often later. Although feedback is considered essential to learning, what is it and how can tutors be supported to provide pertinent feedback to their students?

Feedback is, put simply; additional tutoring that is tailored to the learner's current needs. In the simplest case, this means that there is a mismatch

between students' and the tutors' conceptual models, and the feedback is reducing or correcting this mismatch, very much as feedback is used in cybernetic systems. This is not an accident, for the cybernetic analogy was based on Pask's (1976) work, which has been a strong influence on practice in this area (e.g., Laurillard, 1993).

Because feedback is very much at the cutting edge of personal learning, we wanted to see how we could work with tutors to improve the quality of their feedback. To achieve this, we have been working on tools to provide tutors with opportunities to reflect on their feedback. The latest of these, Open Mentor, is an open source tool which tutors can use to analyse, visualise, and compare their use of feedback.

In particular, we wanted to consider feedback not as error correction, but as part of the dialogue between student and tutor. This is important for several reasons: first, thinking of students as making errors is unhelpful – as Norman (1988) says, errors are better thought of as approximations to correct action. Thinking of the student as making mistakes may lead to a more negative perception of their behaviour than is appropriate. Secondly, learners actually need to test out the boundaries of their knowledge in a safe environment, where their predictions may not be correct, without expecting to be penalised for it. Finally, feedback does not really imply guidance (i.e., planning for the future) and we wanted to incorporate that type of support without resorting to the rather clunky 'feed-forward'.

In this paper, we will describe Open Mentor, and the processes that we worked through as we developed it. We started the process by checking with our stakeholders, i.e. tutors and students, that our tutoring model was one they recognised and welcomed.

## **Background**

In order to provide feedback, Open Mentor has to analyse the tutor comments.

The classification system used in Open Mentor is based on that of Bales (1970). Bales's system was originally devised to study social interaction, especially in collaborating teams; its strength is that it brings out the socio-emotive aspects of dialogue as well as the domain level. In previous work (Whitelock et al., 2004) we found that the distribution of comments within these categories correlates very closely with the grade assigned.

Bales' model provides four main categories of interaction: positive reactions, negative reactions, questions, and answers. These interactional categories illustrate the balance of socio-emotional comments that support the student. We found (Whitelock et al., 2004) that tutors use different types of questions in different ways, both to stimulate reflection, and to point out, in a supportive way, that there are problems with parts of an essay. These results showed that about half of Bales's interaction categories strongly correlated with grade of assessment in different ways, while others were rarely used in feedback to

learners. This evidence of systematic connections between different types of tutor comments and level of attainment in assessment was the platform for the current work.

The advantage of the Bales model is that the classes used are domain-independent – we used this model to classify feedback in a range of different academic disciplines, and it has proven successful in all of them. An automatic classification system, therefore, can be used in all fields, without needing a new set of example comments and training for each different discipline.

Others (e.g., Brown & Glover, 2006) have looked at different classification systems, including Bales, and from these developed their own to bring out additional aspects of the tutor feedback, bringing back elements of the domain. In practice, no (useful) classification system can incorporate all comments. We selected, and still prefer, Bales because of its relative simplicity, its intuitive grasp by both students and tutors, and because it brings out the socio-emotive aspects of the dialogue, which is the one aspect tutors are often unaware of.

A second point is that Bales draws out a wider context: we found that as we started to write tools that supported feedback, we began to question the notion of feedback itself. Instead, the concept seemed to divide naturally into two different aspects: learning support and learning guidance. Support encourages and motivates the learner, guidance shows them ways of dealing with particular problems.

### **Understanding the stakeholders needs**

In order to build the first storyboards for Open Mentor and to ensure the software would meet the needs of both tutors and students, we devised two questionnaires, one for tutors and the other for students. 44 tutors from Kings College London, Manchester Metropolitan, The Open University and Robert Gordon University completed the tutor questionnaire while 47 students from The Open University and Robert Gordon University responded to a questionnaire which was designed to understand how students reacted to tutor feedback.

The first set of questions raised with both students and tutors perceptions about when written comments on assignments were read by the students. All student respondents indicated that they look at the marks first (rather than comments) and this fitted with the tutors' perceptions.

Most students indicated that they read all comments (Chi Square 12.4  $p < 0.02$ ), while some skimmed comments and few read each point in detail. However, the majority of tutors thought that students mainly skimmed comments (Chi Square 21.636  $p < 0.001$ ) while some did not read them often or read all or in detail. Here, students' responses and tutor perceptions did not agree. In fact they did not believe the students paid as much attention to their feedback as reported by the students.

The vast majority of students reported that they read comments immediately (Chi Square 22.638  $p < 0.001$ ) but never again and this corresponds with tutors' judgments (Chi Square 59.905  $p < 0.001$ ) as to how they thought students behaved. However 19 students reported that they later refer back to comments, an observation that is not reflected in the tutors' judgements.

Both tutors and students agreed that comments should reflect the grade awarded, which is a basic premise of the Open Mentor system.

The majority of tutors involved in the study judged themselves to be experienced tutors. The majority of student respondents did not judge there to be a difference in feedback from new and experienced tutors. However, some students reported that new tutors provided more feedback than experienced tutors.

The majority of tutors indicated that new tutors provide students with the greatest amount of written feedback while a significant number felt that there was no difference between tutors. With respect to the quality of feedback however, the majority of tutors felt that experienced tutors provided higher quality (Chi Square 10.878  $p < 0.004$ ) while a significant number felt that there was no difference between experienced tutors and others. This result is the opposite of student judgements where a majority felt that there was no difference between new and experienced tutors, while a significant number (Chi Square = 19.0  $p < 0.01$ ) thought that experienced tutors provided better comments.

A large majority of tutors and students indicated that a software tool to assist with commenting would be of help to tutors and in training tutors. All tutors felt that software tool would help them reflect upon feedback to students but tutors were divided about how such a tool might help with the management of resources. However, a significant majority of tutors felt that a software tool would be of help with Quality Assurance (Chi Square = 18  $p < 0.01$ )

Responses to open ended questions were very diverse among both students and tutors. However, both groups indicated that students most value constructive positive comments even if critical. Similarly both groups felt that there is little value in negative comments and unsubstantiated comments. Both groups indicated that feedback should be improved through more detail and that comments should be meaningful, constructive and relate to the actual assessment. Finally, there were consistent comments that experienced tutors have a better understanding of students while new tutors might be more enthusiastic.

#### *Questions which tested the underlying pedagogical model for Open Mentor*

Previous work by Whitelock, Watt, Raw and Moreale (2004) on student feedback has postulated that work that is awarded high grades should attract feedback from tutors that is high in praise, has few questions and does not ask the student to reflect on their work. Conversely, work that is awarded low grades should attract less praise, more questions and suggestions and invite



more reflection. A number of questions in the Open Mentor Evaluation Study are able to throw light on these postulated outcomes and the results are summarised below.

A significant majority of both students and tutors respondents indicated that they expected high grades to attract more positive comments and low grades to attract more answers, suggestions and questions. Tutors gave a strong indication that they expected assessments with low grades to attract negative comments. Student responses followed a similar trend that was however not statistically significant. Students also indicated strongly that they expected no difference. All these findings support the pedagogical model postulated by Whitelock et al.

A further analysis, using cross tabulation revealed:

- Both students and tutors who feel that low grades would result in more questions also indicated that low grades would attract more answers
- Tutors who judged that high grades attract more positive comments also indicated strongly that low grades attract more answers and suggestions
- Tutors who felt that low grades attract more questions also indicated that low grades attract negative comments
- Both students and tutors felt that lower grades should attract more detailed comments and a deeper level of explanation. Higher grades should attract more positive comments

These findings from both groups of stakeholders supported a pedagogically driven development process which is described below.

### **The design of Open Mentor**

We followed a process that began with developing scenarios of use, then storyboards, and then putting in place an implementation which would follow closely the pattern of these storyboards.

The idea behind the design of Open Mentor is fairly straightforward: it goes through tutor assignments, extracting tutor comments, and classifying them. We used pre-determined benchmarks (from Whitelock et al., 2004, although these can be adapted to different institutions) to estimate 'ideal' distributions of comments for each category, and then display the difference between the actual and the ideal. In practice, this is a bit of a simplification – the actual logic is pretty complex, but most of this is hidden. Although there are 'normal' bands of comments of each type, these vary (significantly) depending on the quality of the individual submissions and the number of submissions involved. A large proportion of positive comments in one context may be inappropriate in a second, and coincidental in a third.

Open source was initially an external requirement, but subsequently became a way of life. The two rounds of the project were funded by JISC, which mandated open source where possible. Initially, this meant re-using other people's code where we could, basically to save us having to do the work ourselves. Ultimately, though, open source changed the way we designed the system into a far more open structure than we had initially conceived.

The resulting Open Mentor architecture is based on the following main components:

- A data source for course information and lists of students and tutors
- A data source for use within Open Mentor, to store assignments, submissions and classified comments
- A classifier which can categorise tutor comments
- An extractor which can read tutor comments from word processed files
- An evaluation scheme description which defines the classes of comments, the grading bands and the expected benchmarks
- A logic component which applies the evaluation scheme to the classified comments

The advantage of this is that different institutions can write their own components and add them into the system without having to do any modification of existing code – this reduces the risk of errors and other problems.

### **How does Open Mentor work?**

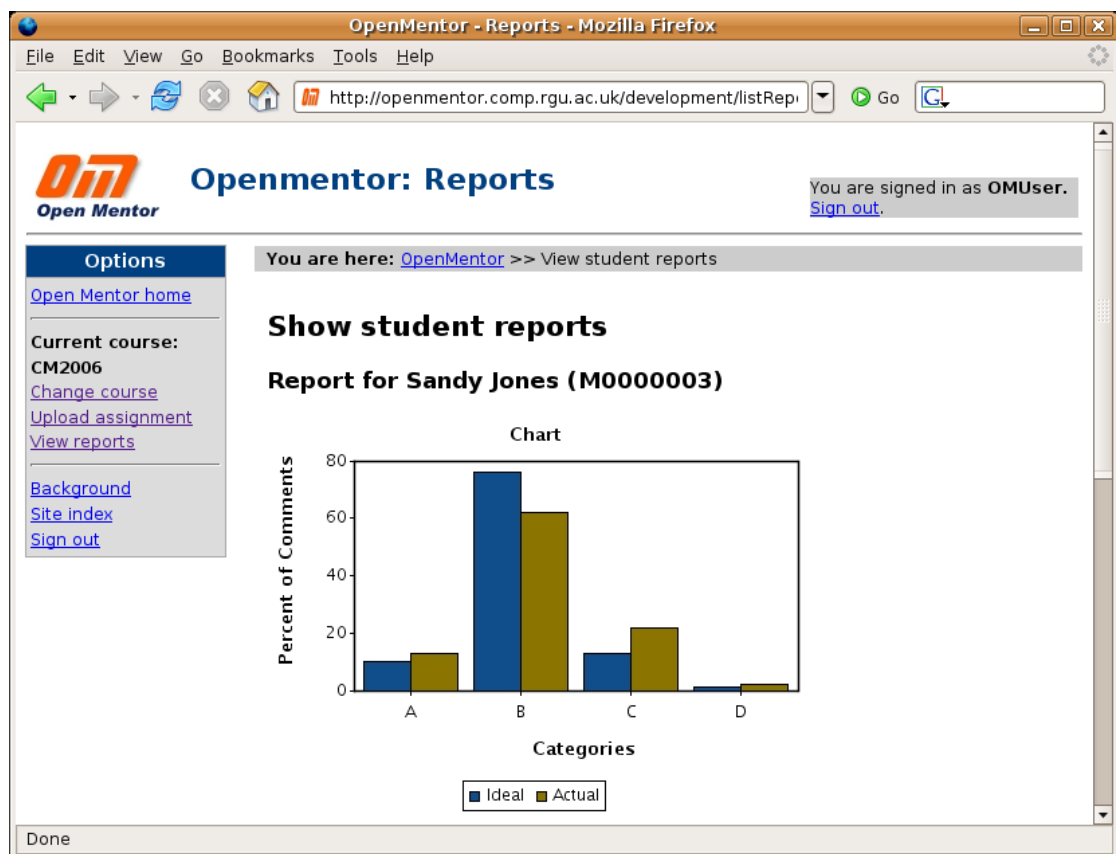
Open Mentor reads and opens assignments written in Microsoft Word to extract the tutor comments. However, it does not use Word itself. A standard charting component is used to provide interactive bar chart views onto the tutors' comments showing the difference between actual and ideal comment distributions as shown in Figure 1 below. It provides the tutor with feedback about the types of comments s/he has given to the student with respect to the mark awarded. If there is not enough praise or constructive feedback for improvement then the tutor will be alerted to this finding.

### **The implementation of Open Mentor**

Open Mentor is implemented using Java, and runs as a web application, enabling people to use it in any location. A screen snapshot of the system is shown below in figure 1.

Again, open source influenced the direction of the project; initially we had used open source as a kind of library of components that we could re-use. Later, particularly when we moved to Spring, we found our system became much smaller, as we could plug our developments more easily into larger

frameworks. We also moved to a point where we can contribute to open source: our developments of the Apache POI-based code for extracting text from Word files exceeded the capabilities of the standard distribution. UK higher education has an important new dissemination route for its developments in these channels – however, the same resourcing issues that led to this situation still need to be addressed.



**Figure 1: Screen snapshot of Open Mentor**

In our original implementation, we used Open Office to read Microsoft Word files, converting them to the Open Office (then) standard .sxw files, which in practice are zipped XML files. Open Mentor would ask Open Office to convert the file, then it would unzip the resulting .sxw file, parse the XML, and extract the tutor comments. Although this worked, it made the server dependent on Open Office, which proved to be too unstable for a reliable system. The current version of Open Mentor uses a separate comment extraction component, based on Apache's POI (a library for working with Microsoft OLE-based documents, especially Excel). This means that Open Office does not need to be running, and everything can be managed within a single Java application, improving reliability considerably.

The currently implementation uses the Spring framework to divide the system into a larger body of components, each of which can be used, replaced, or wrapped independently, making the overall system easier to integrate into an

existing framework. Each of the blocks in the diagram above were represented by one or more components.

Other than this, the implementation is a fairly standard Java-based open source framework. We used JSTL to implement the web pages, and Hibernate to map business objects into a relational database. We also used a few other tools to support the method, and especially, we used Maven – an advanced Java build management tool, which enabled us to track the quality of the development work in a distributed team. Subversion proved a great alternative to CVS which worked well through our somewhat complex firewall arrangements.

## **Discussion**

What of the connection between education and technology? As a development team we were fortunate, as many of the team combined both technical expertise and long experience of innovation in pedagogy. This enabled us to let the needs of the tutors and students drive the technology. In more traditionally structured teams this would have been either impossible or frustrating, and quite likely both – as control within teams flows between technological and educational specialists. To be successful, we had to become specialists in both.

Open Mentor is becoming successful, both within our institutions and beyond. However, the key factor is still institutional integration, and the key to this is the open frameworks that are enabled by the use of open source. In today's educational climate, with the continued pressure on staff resources, making individual learning work is always going to be a challenge. But it is achievable, so long as we manage to maintain our empathy with the learner. Tools can help us achieve this by giving us frameworks where we can reflect on our social interaction, and ensure that it provides the emotional support as well as the conceptual guidance that our learners need.

## **Acknowledgements**

We wish to thank Colin Beagrie, Jan Rae and Jan Holt for their support with project management, data collection and analysis during this project and to the JISC for supporting this type of software development.

## References

Bales, R.F (1950) A set of categories for the analysis of small group interaction. *American Sociological Review*, 15:257-63

Beck, K. (2002). The metaphor metaphor. Presented at OOPSLA'2002, Portland, Oregon, October 22<sup>nd</sup> to 26<sup>th</sup>.

Brown, E., & Glover, C. (2006). Evaluating written feedback. In *Innovative Assessment in Higher Education*, (eds., Bryan, C., & Clegg, K.), Routledge, pp. 81-91.

Brown, E., & Glover, C. (2006). Written feedback for students: too much, too detailed or too incomprehensible to be effective? *Bioscience Education e-Journal*, 7(3).

Laurillard, D. (1993). *Rethinking University Teaching: A Framework for the Effective Use of Educational Technology*. London: Routledge.

Norman, D. (1988). *The psychology of everyday things*. New York: Basic Books.

Pask, G. (1976). *Conversation theory: applications in education and epistemology*. Amsterdam: Elsevier.

Rosson, M. B., & Carroll, J. M. (2002). *Usability engineering: scenario-based development of human computer interaction*. San Francisco: Morgan Kaufmann.

Whitelock, D., Watt, S. N. K., Raw, Y., & Moreale, E. (2004). Analysing tutor feedback to students: first steps towards constructing an electronic monitoring system. *ALT-J*, 1(3), 31-42.

Whitelock, D. and Brasher, A. (2006). Developing a Roadmap for e-assessment: which way now? CAA Conference 2006, Loughborough University, 4/5 July 2006.



# **USING DIGITAL STORYTELLING AS AN ASSESSMENT INSTRUMENT: PRELIMINARY FINDINGS AT AN ONLINE UNIVERSITY**

**Jeremy B. Williams and Kanishka Bedi**





# Using Digital Storytelling as an Assessment Instrument: Preliminary Findings at an Online University

Jeremy B. Williams and Kanishka Bedi

Universitas 21Global

[jwilliams@u21global.edu.sg](mailto:jwilliams@u21global.edu.sg)

[kbedi@u21global.edu.sg](mailto:kbedi@u21global.edu.sg)

## Introduction

'Digital Storytelling' is a term often used to refer to a number of different types of digital narrative including web-based stories, hypertexts, videoblogs and computer games. While the definition of digital storytelling is still evolving, this emergent form of creative work has found an outlet in a wide variety of different domains ranging from community social history, to cookbooks, to the classroom. It is the latter domain that provides the focus for this paper, specifically the online classroom in the graduate business school environment.

The authors hypothesise that as – in the majority of societies – people are 'hard wired' both to tell and to listen to stories from a very young age and, significantly, to *remember* stories, the scope for deep learning using this particular pedagogical tool is considerable. The more conservative forces within business schools may not be persuaded by this idea but – whether they are or not – the fact remains that, in the knowledge economy, digital technologies have become the *modus operandi* for business communication. In this sense, a business school curriculum with a heavy bias towards text-based, essay-style assignments might be adjudged out-of-step with the times. A supplementary hypothesis, therefore, is that digital storytelling also represents a highly authentic form of assessment (Herrington et al. 2003), in that the digital storytelling format improves presentation skills which are highly sought in the business world today.

Much of the work on digital storytelling in the education sphere has concentrated on the primary and secondary sectors. With some notable exceptions (e.g. Paull 2002), the literature on digital storytelling in the tertiary/adult education sector is quite sparse. Research on the use of digital storytelling in business schools, meanwhile, appears non-existent, hence the motivation for this study.

## Methodology

In early 2006, work began at Universitas 21 Global (U21Global) – a completely online university – to investigate the extent to which digital storytelling might be integrated into the MBA course. In keeping with the work of Paull (2002), the initial focus was how digital stories might help to create a sense of personal and social agency and empowerment within students; characteristics that appeared to be in short supply within U21Global's virtual distance learners. This involved training faculty in the art of digital story creation in order to better introduce themselves to their students and, thereafter, have students reciprocate by producing their own stories (see Williams, Bedi and Goldberg, 2006). Current work-in-progress has concentrated on the use of digital storytelling as the vehicle for the submission of assignments.

In late 2006, an experiment was conducted with an Operations Management class where participants were required to submit a team assignment and a team-based final project in digital story format using Microsoft PowerPoint. As a standard inclusion in the Microsoft Office suite, this application is used extensively by faculty and students in universities the world over, but the use of images, animation and voice is quite rare in the traditional business school setting; there being a greater tendency to rely on text. This break with tradition enables the student to become more *actively* involved in their learning and not simply rehash and regurgitate text cut-and-pasted from various electronic resources. Importantly, it permits the student to *construct* a more authentic and meaningful learning context for themselves in which cases are brought to life through real-world images (or through images used as metaphor), and through the use of their own narrative rather than somebody else's for whom they may have little real empathy.

At the end of the course, the students were invited to take part in an 18-question semi-structured survey covering a number of different aspects of digital storytelling.

## Findings of the 2006 pilot study

While the results of the experiment are only preliminary at this stage (given it was a pilot study), positive feedback from 24 respondents to the questionnaire (a response rate of 69% in a class of 35) suggests that listening to and telling 'true stories' was a compelling and engaging experience, providing an opportunity for 'transformative reflection' (Lambert 2000). By including multimedia, learners were able to build upon the fundamentals, presenting content in an easy-to-absorb and compelling way. In terms of team assignments students learned to become more effective actors in collaborative work environments, and felt encouraged to communicate meaning on multiple levels. Importantly, this approach offered an entertaining way of promoting team awareness and coherence in virtual teams. It also provided an avenue for the students in this class to express their creativity. Some of them came up with their own unique methods of creating digital

stories using Flash and Camtasia, rather than use the narrated PowerPoint model suggested by faculty. On the minus side, not all the respondents were equally enamoured by the digital storytelling experience, and there would appear to be a number of obstacles in the path of a programme-wide roll out of digital storytelling as a formal assessment instrument.

The moderate success of digital storytelling as an assessment tool in this pilot study has been followed up with further experimentation in other MBA classes during early 2007, with a view to addressing some of the problems identified by the detractors. One important improvement has been to provide additional scaffolds for learners who are enthusiastic digital storytellers, but feel challenged by the *process* of compiling a digital story. In the 2006 pilot study, students were provided with a digital story created by faculty entitled “Creating Digital Stories”, the aim of which was to provide some basic instructions and to reduce the ‘learning curve’ for those less adept at using PowerPoint. It was also thought sufficient for faculty to act as role model for the students in the creation of their personal introduction in digital storytelling format, and that this format serve as a template (or at least a guide) for the assignments that their students would submit later that term. In summary, an assumption was made that, as students at an online university, learners would possess sufficient ‘tech-savviness’ to put together a narrated PowerPoint without too much difficulty. This assumption was flawed on two counts; (i) that all students would have sufficient technical expertise, and (ii) that – technical considerations aside – the principles of digital storytelling (as they were articulated in the context of a personal introduction) would be seamlessly applied in the context of an assignment for formal assessment purposes. Thus, in the 2007 study, learners were provided with two additional scaffolds; an exemplar digital story team assignment from a previous class, and a “Digital Story FAQ” presentation (in digital story format) created by faculty to specifically address the common technical problems encountered and the key principles one might adhere to in composing a digital story for assessment purposes.

### **Findings of 2007 follow-up study and comparative analysis**

The 2007 survey tool (see Appendix) was the same as that used in the 2006 pilot study with a few modifications. Two questions were added to take account of the additional scaffolds described above, and two questions were omitted; one focusing on the inclusion of digital stories in discussion board postings, a practice that was discontinued after the pilot project; and one question referring to the use of the digital story format in the students’ Final Project (a practice that was also discontinued). The same Operations Management course was the vehicle for the follow-up study and in a class of 29 students there were 22 responses (a response rate of 76%).

In both the 2006 and 2007 surveys the download and viewing of the digital story introduction of the professor was considered a straightforward process by the majority of students. Furthermore, all either agreed or strongly agreed that the professor’s digital story helped them to get to know him/ her better

compared to the usual text-based introduction. In the 2006 survey, the majority of the students felt that the professor's digital story 'improved the learning environment' for them – the 2007 result being slightly less resounding – the comments of students suggesting that the learning environment does tend to be very impersonal, so the digital story introduction helped them to make a connection with the professor. Interestingly, however, the quantitative data did not reflect the sentiment contained within the qualitative data, as in both the surveys, only 40% of the students felt that digital storytelling should be a feature of all U21Global subjects for introductions by professors and students; a large majority remaining neutral.

On the question of whether the opportunity to submit the Team Assignment in digital storytelling format was a good idea, the 2006 survey received a more favourable response from the students compared to the 2007 survey. This was a disappointing result, especially given students had commented in response to the 2006 survey that it would be better if the digital story format were restricted to only the Team Assignment, rather than both the Team Assignment and the Final Project. With hindsight, however, it is probably not appropriate to compare the responses to this question as the exposure to digital storytelling of each class is different. It could be, for example, that the responses of the 2007 class would have been different had they also submitted their Final Projects in digital story format, or that the 2006 class would have thought differently had they only done the Team Assignment as a digital story. Whichever way one interprets this result, there would appear to be sufficient doubt over the veracity of the questionnaire in this instance, for this issue to warrant further study. This is especially the case when one considers the qualitative data in relation to the use of the digital story format for assignment submission which corroborate its usefulness in terms of the skill development it facilitates; for example:

*“This compensated for the missing opportunity for U21 students to do oral presentation which normally happens in other MBA courses.”*

*“This has also helped us to summarize our report and be more creative.”*

*“Really got our creative juices flowing.”*

In the part of the survey instrument devoted to understanding the experiences of students in creating the digital story assignment, a majority in both surveys responded that deciding what information to include was a straightforward process. So, too, was the process for finding relevant images from the Web for compilation of the digital story. However, while recording the narration was straightforward for 2007 class (more than 82% students agreeing or strongly agreeing), this was not the case for the 2006 class where only 37% agreed or strongly agreed. This is a pleasing result as it would suggest that the additional scaffolding included for the 2007 class paid off; the “Digital Story FAQs” presentation and the sample digital story assignment serving to increase the comfort level for students when recording their narrations. Even so, students' comments still indicate that this is a difficult skill to acquire:

*"For students who are familiar with presentation in ppt and who are knowledgeable of the information required for presentation, it is quite a breeze. However, recording seemed to "magnify" mistakes and "ah.ah.eh.eh", and for a perfectionist, it can take quite a bit of efforts in re-recording! But great training for presentation skills!"*

*"It did take a long time for the initial recording as well as getting the timing correct. I did experience that it took many "takes" to get it right, but the lesson learned - prepare a proper script to eliminate hesitation and make the presentation flow."*

Another positive outcome to emerge from the 2007 data was that the difficulty encountered by students in uploading the digital story assignment file seemed to have decreased. Less than 20% of the students indicated they experienced such problems in the 2007 class compared to more than 40% in the 2006 class. This improvement can be largely attributed to the provision of a dedicated student FTP site for the 2007 class, while the earlier class had to upload via a Learning Management System less able to cope with large file sizes. In keeping with this result, the proportion of students feeling that, overall, the creation of the digital story for assignment submission was a relatively straightforward process, increased from about 45% in 2006 to 60% in 2007. Further, in the 2007 class, about 80% of the students were of the view that the "Creating Digital Stories: Principles and Practice" presentation, the "Digital Story FAQs" presentation and the sample digital story assignment were helpful in creating the digital story team assignment.

In both surveys, more than 55% of the students agreed that the submission of the assignment(s) in digital storytelling format improved the learning outcomes from the subject. While this is a satisfactory result, around one third surveyed disagreed. More encouraging was that in the 2006 survey, 75% of the students felt that the submission of the assignment(s) in digital storytelling format improved this type of presentation skill which is highly sought in the business world today. In the 2007 survey, 82% of the students believed this to be the case. The following comments provide an illustration of this strong endorsement:

*"This was good practice for me and more like a real business assignment than the written assignments."*

*"This is exactly the skill missing in an online MBA course. The ability to present well by being able to communicate in person and to "project" the energy and enthusiasm of the idea is important if not the most important skill senior managers need for internal and external selling to top management members in order to succeed."*

*"More visual !!!! and effective !!!!"*

*"Digital story telling (when used in office environment) makes it superior form of ppt as compared to the standard presentation techniques."*

*“With the current work environment, no senior managers would like to spend the time or want to spend the time to pour through pages and pages of text written document, As I have discovered that after having written 20 pages of strategy on business improvement, my senior GM refuse to read it , so I had to condensed it into a ppt file where it will display the key points. It is time to keep up with the changes in the real business world. After the ppt presentation, my senior GM complemented me on an excellent piece of work.”*

A common negative theme to emerge in both surveys was the time it took to create digital stories. When asked if they would feel comfortable submitting their OBOW (Open Book Open Web) exam (see Williams 2006) in this format, a majority of students responded in the negative. The concerns expressed by the students with regard to this were primarily based upon the time constraints and possibility of technical glitches taking place during the examination period. Similarly, in the 2006 class, the survey revealed that a majority of students (more than 50%) felt that while the discussion board postings of some students in their class in digital storytelling format were more engaging compared to the usual text-based postings, this practice should not become mandatory as time would be a major constraint.

In 2006 as well as the 2007 survey, a majority of students (more than 50%) either agreed or strongly agreed that digital storytelling should be a feature of all U21Global subjects for at least one assignment. However, a larger number of students (36%) were against it in the 2007 class compared to about 8% against it in the 2006 class.

## **Conclusions and future directions**

Further validation is clearly required before contemplating large-scale implementation and this is unlikely to be realised until several more studies have been completed in different discipline areas within the MBA course. A significant challenge yet to be tested is the resistance (or otherwise) of faculty to the widespread adoption of digital storytelling as a reliable and valid assessment instrument.

These challenges aside, the proliferation of broad-band Internet access and the increasing availability of file compression software have opened up new exciting vistas in higher education. It may take some time before the concept of digital storytelling takes hold, but as this paper suggests, existing obstacles are not insurmountable, and with further experimentation and analysis, digital storytelling has the potential to become a mainstream assessment instrument even in the traditionally conservative environment of a graduate business school.

## References

Herrington, J., Oliver, R., & Reeves, T. C. (2003). Patterns of engagement in authentic online learning environments. *Australian Journal of Educational Technology*, 19(1), 59-71.

Paull, C. (2002). *Self-Perceptions and Social Connections: Empowerment through Digital Storytelling in Adult Education*, University of California, Berkeley, Dissertation Abstracts International.

Lambert, J. (2000). Has digital storytelling succeeded as a Movement? Some thoughts. *dStory News*, Issue 2, September 20. Available online: [http://www.dstory.com/dsf6/newsletter\\_02.html](http://www.dstory.com/dsf6/newsletter_02.html) (24 February 2007).

Williams, J.B., Bedi, K. & Goldberg, M.A. (2006). The impact of digital storytelling on social agency: early experience at an online university. U21Global Working Paper Series, 003/2006, August. Available online: <http://u21global.com/PartnerAdmin/ViewContent?module=DOCUMENTLIBRARY&oid=157295> (24 February 2007).

Williams, J.B. (2006) 'The place of the closed book, invigilated final examination in a knowledge economy', *Educational Media International*, 43(2), 107-119.

## Appendix

### 2007 Questionnaire

**1. The introduction of the professor in digital story format helped me to get to know him/ her better compared to the usual text-based introduction.**

- ☐ 1. Strongly Disagree
  - ☐ 2. Disagree
  - ☐ 3. Neutral
  - ☐ 4. Agree
  - ☐ 5. Strongly Agree
  - ☐ Not applicable
- 

**Please add any other comment you feel is relevant:**

Not Applicable

---

**2. The introduction of the professor in digital story format has improved the learning environment for me.**

- ☐ 1. Strongly Disagree
  - ☐ 2. Disagree
  - ☐ 3. Neutral
  - ☐ 4. Agree
  - ☐ 5. Strongly Agree
  - ☐ Not applicable
- 

**Please add any other comment you feel is relevant:**

Not Applicable

---

**3. The opportunity to submit the Team Assignment in digital storytelling format was a good idea.**

- ☐ 1. Strongly Disagree
- ☐ 2. Disagree



- ☐ 3. Neutral
  - ☐ 4. Agree
  - ☐ 5. Strongly Agree
  - ☐ Not applicable
- 

**Please add any other comment you feel is relevant:**

Not Applicable

---

**4. Deciding what information to include in the digital story was a straightforward process.**

- ☐ 1. Strongly Disagree
  - ☐ 2. Disagree
  - ☐ 3. Neutral
  - ☐ 4. Agree
  - ☐ 5. Strongly Agree
  - ☐ Not applicable
- 

**Please add any other comment you feel is relevant:**

Not Applicable

---

**5. The submission of the Team Assignment in digital storytelling format improved the learning outcomes from this subject.**

- ☐ 1. Strongly Disagree
  - ☐ 2. Disagree
  - ☐ 3. Neutral
  - ☐ 4. Agree
  - ☐ 5. Strongly Agree
  - ☐ Not applicable
- 

**Please add any other comment you feel is relevant:**

Not Applicable

---

**6. The submission of the Team Assignment in digital storytelling format improved this type of presentation skill, which are highly sought in the business world today.**

- ☐ 1. Strongly Disagree
- ☐ 2. Disagree
- ☐ 3. Neutral
- ☐ 4. Agree
- ☐ 5. Strongly Agree
- ☐ Not applicable

---

**Please add any other comment you feel is relevant:**

Not Applicable

---

**7. I was integrally involved in the creation of digital story for my Team Assignment.**

- ☐ 1. Strongly Disagree
- ☐ 2. Disagree
- ☐ 3. Neutral
- ☐ 4. Agree
- ☐ 5. Strongly Agree
- ☐ Not applicable

---

**Please add any other comment you feel is relevant:**

Not Applicable

---

**8. I would feel comfortable submitting my OBOW (Open Book Open Web) exam in this format.**

- ☐ 1. Strongly Disagree
- ☐ 2. Disagree
- ☐ 3. Neutral
- ☐ 4. Agree
- ☐ 5. Strongly Agree

☐ Not applicable

---

**Please add any other comment you feel is relevant:**

Not Applicable

---

**9. The download and viewing of the digital stories of the Professor was a straightforward process.**

- ☐ 1. Strongly Disagree
  - ☐ 2. Disagree
  - ☐ 3. Neutral
  - ☐ 4. Agree
  - ☐ 5. Strongly Agree
  - ☐ Not applicable
- 

**Please add any other comment you feel is relevant:**

Not Applicable

---

**10. Creating the MS PowerPoint slides for the digital story was a straightforward process.**

- ☐ 1. Strongly Disagree
  - ☐ 2. Disagree
  - ☐ 3. Neutral
  - ☐ 4. Agree
  - ☐ 5. Strongly Agree
  - ☐ Not applicable
- 

**Please add any other comment you feel is relevant:**

Not Applicable

---

**11. Finding the relevant images from the web for the digital story was a straightforward process.**

- ☐ 1. Strongly Disagree
  - ☐ 2. Disagree
  - ☐ 3. Neutral
  - ☐ 4. Agree
  - ☐ 5. Strongly Agree
  - ☐ Not applicable
- 

**Please add any other comment you feel is relevant:**

Not Applicable

---

**12. Recording the narration in the background for the digital story was a straightforward process.**

- ☐ 1. Strongly Disagree
  - ☐ 2. Disagree
  - ☐ 3. Neutral
  - ☐ 4. Agree
  - ☐ 5. Strongly Agree
  - ☐ Not applicable
- 

**Please add any other comment you feel is relevant:**

Not Applicable

---

**13. The file size of the completed digital story meant uploading via the “Student upload site” and this was manageable.**

- ☐ 1. Strongly Disagree
  - ☐ 2. Disagree
  - ☐ 3. Neutral
  - ☐ 4. Agree
  - ☐ 5. Strongly Agree
  - ☐ Not applicable
- 

**Please add any other comment you feel is relevant:**

Not Applicable

---

**14. Overall, the creation of the digital story for assignment submission was a relatively straightforward process.**

- ☐ 1. Strongly Disagree
  - ☐ 2. Disagree
  - ☐ 3. Neutral
  - ☐ 4. Agree
  - ☐ 5. Strongly Agree
  - ☐ Not applicable
- 

**Please add any other comment you feel is relevant:**

Not Applicable

---

**15. The “Creating Digital Stories: Principles and Practice” presentation was helpful in creating the digital story team assignment.**

- ☐ 1. Strongly Disagree
  - ☐ 2. Disagree
  - ☐ 3. Neutral
  - ☐ 4. Agree
  - ☐ 5. Strongly Agree
  - ☐ Not applicable
- 

**Please add any other comment you feel is relevant:**

Not Applicable

---

**16. The “Digital Story FAQs” presentation was helpful in creating the digital story team assignment.**

- ☐ 1. Strongly Disagree
  - ☐ 2. Disagree
  - ☐ 3. Neutral
  - ☐ 4. Agree
  - ☐ 5. Strongly Agree
  - ☐ Not applicable
- 

**Please add any other comment you feel is relevant:**

Not Applicable

---

**17. The sample digital story assignment provided was helpful in the creation of digital story team assignment.**

- ☐ 1. Strongly Disagree
  - ☐ 2. Disagree
  - ☐ 3. Neutral
  - ☐ 4. Agree
  - ☐ 5. Strongly Agree
  - ☐ Not applicable
- 

**Please add any other comment you feel is relevant:**

Not Applicable

---

**18. Digital storytelling should be a feature of all U21Global subjects for introductions by professors and students.**

- ☐ 1. Strongly Disagree
  - ☐ 2. Disagree
  - ☐ 3. Neutral
  - ☐ 4. Agree
  - ☐ 5. Strongly Agree
  - ☐ Not applicable
- 

**Please add any other comment you feel is relevant:**

Not Applicable

---

**19. Digital storytelling should be a feature of all U21Global subjects for at least one assignment.**

- ☐ 1. Strongly Disagree
  - ☐ 2. Disagree
  - ☐ 3. Neutral
  - ☐ 4. Agree
  - ☐ 5. Strongly Agree
  - ☐ Not applicable
- 

**Please add any other comment you feel is relevant:**

Not Applicable

**20. I have some suggestions regarding the use of digital storytelling in assignment submission as under:**

Not Applicable





# **AN E-LEARNING FRAMEWORK FOR ASSESSMENT (FREMA)**

**Gary B. Wills<sup>1</sup>, Christopher P. Bailey<sup>1</sup>, Hugh C. Davis<sup>1</sup>, Lester Gilbert<sup>1</sup>, Yvonne Howard<sup>1</sup>, Steve Jeyes<sup>2</sup>, David E. Millard<sup>1</sup>, Joseph Price<sup>1</sup>, Niall Sclater<sup>3</sup>, Robert Sherratt<sup>2</sup>, Iain Tulloch<sup>3</sup>, Rowin Young<sup>3</sup>**



# **An e-Learning Framework for Assessment (FREMA)**

Gary B. Wills<sup>1</sup>, Christopher P. Bailey<sup>1</sup>, Hugh C. Davis<sup>1</sup>, Lester Gilbert<sup>1</sup>, Yvonne Howard<sup>1</sup>, Steve Jeyes<sup>2</sup>, David E. Millard<sup>1</sup>, Joseph Price<sup>1</sup>, Niall Sclater<sup>3</sup>, Robert Sherratt<sup>2</sup>, Iain Tulloch<sup>3</sup>, Rowin Young<sup>3</sup>

1 University of Southampton,

2 University of Hull,

3 University of Strathclyde,

## **Abstract**

The paper reports on the FREMA (Framework Reference Model for Assessment) project that aims at creating a Reference Model for the Assessment Domain and delivering it via a heavily interlinked Web site. Because the resulting network of resources (standards, projects, people, organisations, software, services and use cases) is so complex, we require a method of providing users with a structured navigational method that does not require them knowing at first what they might want to find. This led us to look at how overviews of e-learning domains have been handled previously, and work towards our own concept maps that plot the topology of the domain. FREMA was never intended to be a static resource and therefore we converted the original site to use a semantic Wiki, thereby allowing the Assessment Community to use the Knowledgebase to record their own projects, services and potentially new reference models.

## **Introduction**

It is possible to characterise e-learning in terms of a number of domains that group related activities, such as managing e-portfolios or constructing learning content. The assessment domain is one of the most mature of these in terms of software and standards. Numerous commercial and academic tools are available, supporting a wide range of assessment activities, from assembling and running tests or exams to managing feedback and detecting plagiarism.

This raises problems when it comes to building new tools and creating new standards for the domain, as these must be correctly situated with existing work if they are to be successful. This problem is increasingly important in the world of Service-Oriented Architectures (SOA), as new services only become effective when they extend or support existing services. 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.

The service approach is ideally suited to more loosely coupled systems, where individual parts may be developed by different people or organizations. Wilson *et al.* (2004) discuss in detail the advantages of using SOA:

- **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'.

With SOAs there is a need to design complementary services that can be used together to some end. Sometimes these are known as composite services, but in larger cases could represent the infrastructure for an entire domain. Large sets of services that have been designed to work together are often known as service frameworks.

In the UK, the Joint Information Systems Committee (JISC) is financed by all the Further and Higher education funding councils and is responsible for providing advice and guidance on the use of Information and Communications Technology (ICT) for learning and teaching. Part of its strategy is the development of a SOA framework for e-learning (Oliver *et al.*, 2004, Wilson *et al.*, 2004b). JISC call this initiative simply 'e- Framework'.

The e-Framework is based on a service-oriented factoring of a set of distributed core services (Smythe *et al.*, 2004), where flexible granular functional components expose service behaviours accessible to other applications via loosely coupled standards-based interfaces. The technology used is Web Services and the intention is to extend the SOA programming model into a vast networking platform that allows the publication, deployment, and discovery of service applications on the scale of the Internet. However, the e-Framework suffers the same problem as all other service frameworks; mainly that it is difficult to coordinate the development of so many inter-related services by so many people and groups, and disseminate them to the communities that the frameworks serve.

In this paper we present our efforts to develop a Community Reference Model for the development of services within a large Service Oriented framework. Our work has been aimed at the e-Framework, and its development within the domain of e-learning in particular, but the approach is applicable to any service framework that has similar characteristics: i.e. is evolutionary rather than tightly designed, and is being driven forward by distributed, independent developers and users.

## Web services in the assessment domain

In this Section we attempt to give some context to our Reference Model design, by explaining how it is based on concrete problems, faced by real users. In our case this is within the domain of e-learning, and in particular services related to assessment. We show how these real examples, or personas, can be translated into use cases that apply to service frameworks in general.

### *Description of the Assessment Domain*

Conole and Warburton (2005) have recently presented a detailed review of the issues facing computer assisted assessment, and conclude by saying “The role of technology and how it might impact on assessment is still in its infancy and we need to develop new models for exploring this”. Reference models can be thought of as partially filling this need. The e-learning assessment domain has been classified in a number of ways in the past. For instance Bull and McKenna (2004) classify it into four broad categories based on purpose (summative, formative, diagnostic and self-assessment) backed up by a number of taxonomies. JISC themselves have developed a simple map of the assessment domain, using a single test as the connecting thread (Kassam, 2004).

There is a move in learning and teaching to use learning outcomes to define what is to be taught and therefore what is to be assessed. The skill levels defined in the learning outcomes and assessment are often set within Bloom’s (1956) taxonomy of learning objectives. Chang *et al.* (2004) have developed an assessment metadata model (taxonomy) to aid teachers in authoring examinations, which explicitly models the cognition aspect of an assessment in addition to the types of questions.

The e-learning domain is underpinned and sometimes driven by the use of technology. Sclater and Howie (2003) have defined the requirements for the ‘Ultimate’ assessment engine. In presenting these requirements they view the assessment domain from the perspective of the roles people have in the assessment process and how they interact with the resources. Whilst the scope of the assessment domain is open to interpretation, it is likely that core services will include item banks (question databases), delivery applications (that retrieve and render questions) and automatic assessment tools. If the interpretation is broad then services such as peer group formation and plagiarism detection might also be included.

As services within the domain are being developed by a wide variety of institutions for a number of purposes, it is necessary to focus the activities of the assessment community in order that they create interoperable web services and exploit their widest possible use (and re-use). What is required is not just a common repository for services, but a community wide understanding of the domain, and how independently authored services fit within it. If a reference model is to be a community focus point for service design within a framework then it is necessary for it to describe services in the context of well-defined domain processes and also relate them to existing

standards and software. This is a complex challenge due to the many existing e-learning standards, projects, and software.

### *Personas*

We used an agile modelling technique known as 'Personas' in order to investigate the requirements of different members of the assessment community (Cooper *et al.*, 2003). To place personas in a modelling context: if actors and use cases may be considered as abstract classes, then personas and scenarios may be considered instances of those classes where an actor is characterized in detail.

The following are personas that represent the breadth of users that we might expect to interact with a Reference Model:

#### **Persona 1, Will**

'Will is an e-learning tool and web services developer in an academic institution. He is a 30-something post-graduate. He has a good knowledge of the assessment domain and has java and web services technical skills. He is developing an open-source application in the assessment domain focusing on feedback methods.

#### **Scenario:**

*'I want to look up use cases and scenarios to help me design my application. This will help me to define my footprint in the assessment domain. I see there are some web services I could download but some are missing. What standards can I use when writing my own web services to ensure that I can interoperate with the web services I've chosen?'*

#### **Persona 2, Yvonne**

Yvonne is a learning resource manager at a higher education institution with a background in academia and education. She is planning the institution's five year strategy for e-learning. She is responsible for ensuring that new systems meet quality assurance standards. She has a strategic grasp of the importance of e-learning but she is not an expert in the assessment domain.

#### **Scenario:**

*'I want an overview of what this domain is all about. I want to know what standards are applicable in the domain to ensure that we comply with quality assurance requirements. I want to examine use cases and scenarios to understand the available footprints. I also want to know who the key players are and what the key projects are.'*

Although these are just two personas from the assessment community of interest; they have widely different needs and levels of technical expertise and show the range of the spectrum of interaction. Access to resources within the Reference Model should therefore be at different levels of abstraction to match the different characteristics of interest identified.

### *Use Cases*

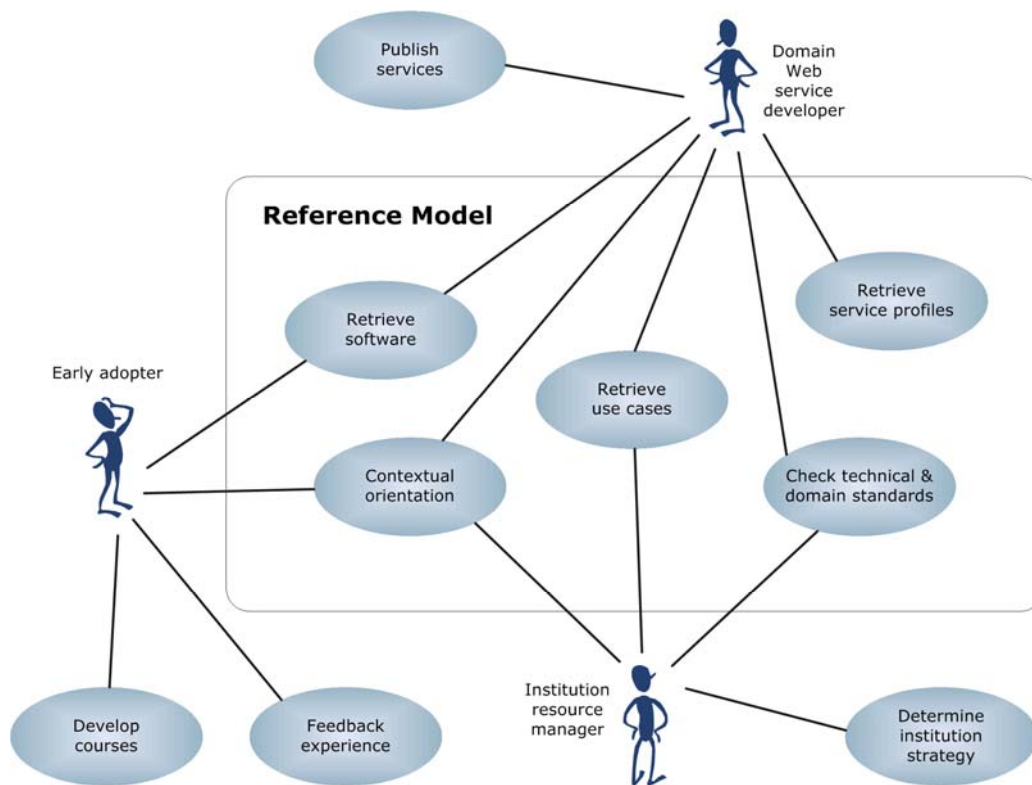
From these and other Personas in the assessment context, we can generalize to three different Reference Model use cases: Domain Web Service Developers, Early Adopters, and Institutional Resource Managers. These are shown together in Figure 1

*Domain Web Service Developers* are actors within the domain that are in the business of creating working software services for a particular framework. They are interested in using the framework to place their own work in the context of the domain (for example, to learn domain vocabulary, and to discover where effort in the domain has been spent), in existing software and standards, and also in domain use cases and service profiles (abstract descriptions of services) that might be related to them.

*Early Adopters* are the primary actors within the domain that want to use emerging technology from the service framework. They are interested in orientating themselves within the domain and also in retrieving software that may help them in their own work.

*Institutional Resource Managers* are actors within the domain who are in charge of institutional policies and direction. They want to use the Service Oriented Reference Model (SORM) to ensure that their institution is using relevant standards in its business processes.

These actors are interested in different technical layers of the reference model. But these layers must be related in order to help the actors orientate themselves and to create an audit trail of decision making throughout the model.



**Figure 1 : Use Cases for a Reference Model**

### **Anatomy of a reference model**

If a Reference Model is to address the needs of such a broad spectrum of users it must contain a wide range of resources, such as descriptions of standards, existing software, use cases, projects, organizations, service profiles, and existing services. However, to be considered a model it must place these in relation with one another, so that it describes the real-world situation. To be an effective model it is necessary for users to be able to understand the model and draw more advantage from it than by examining the real world that is being modelled.

To enable this we have conceptualized a Service Oriented Reference Model as a number of layers, and defined the relationship between each layer. Each layer contains a different set of resources. We have chosen to model these resources ontologically so that the schema of relationships can be shared and understood across the domain. It has also allowed us to create a more dynamic model, which has an extensible set of relation types.

In this section we explain the purpose and content of each layer and describe how, for our Community Reference Model, we exposed the semantic web of resources through a dynamic and heavily interlinked Web site, described at the top level via complementary concept maps.



### *Layered Architecture*

A Service Oriented Reference Model can be thought of as a series of layers. For tightly constrained domains, it may be possible to define a vertical slice through the layers, such that each layer exactly maps onto its vertical neighbours. For broader domains where each layer is smaller in scope but more concrete than the one below it, a Community Reference Model approach is more appropriate.

It is imagined that as a community uses and further develops a Reference Model its higher layers will cover more and more of the lower. Figure 2 shows the layers of the Community Reference Model and the processes that lie between them:

*Domain Definition:* This layer is an overview of the domain that the reference model covers. The definition contains instances from the ontology of domain resources (such as standards, people, and projects) and also the ontological relationships between them. Each of these instances and relationships have narrative descriptions associated with them. In addition each instance is placed in one or more concept networks, so that they may be found by users graphically browsing the domain.

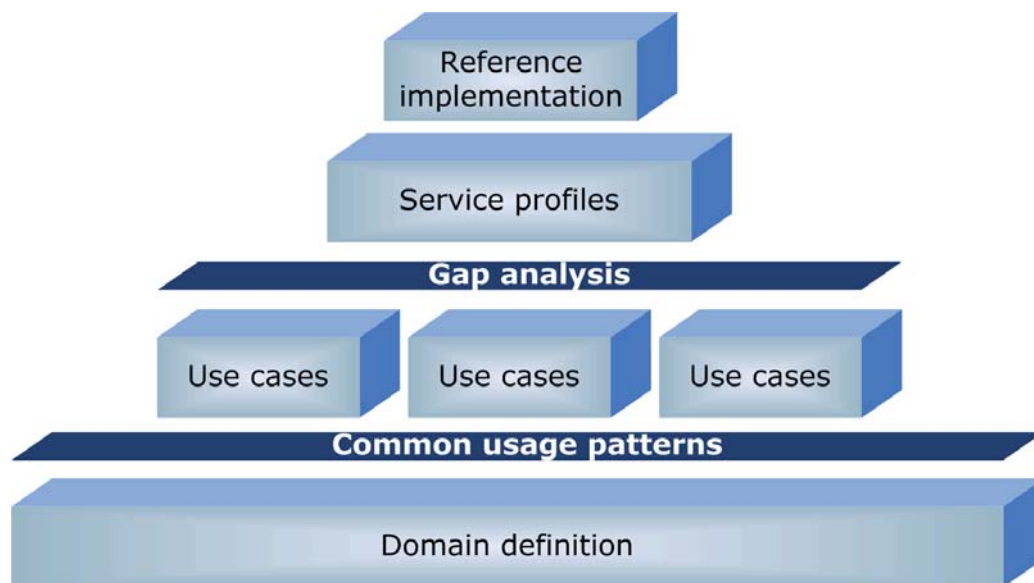
*Identifying Common Usage Patterns:* This is the process of scoping the domain into a manageable subset. Manageable may mean areas that lie unarguably within the domain (according to the views of domain experts), or it may be a reflection of the resources available to create the higher level, more concrete layers of the model. In either case the patterns should include all key activities.

*Use Cases:* This layer formalises the usage patterns into use cases: formal descriptions of user activity in both diagrammatic and narrative form. The Use Cases become new resources, linked to each other and the rest of the Domain Definition through new ontological relationships and narrative descriptions.

*Gap Analysis:* This is the process of mapping the Use Cases to atomic services within a given framework and identifying which ones are missing a formal definition. Not all use cases will necessarily be mapped, although core activities should be covered.

*Service Profiles:* This layer contains the descriptions of those services identified in the gap analysis. *Service Profiles* are abstract descriptions of a service that may be fulfilled by several different *Service Implementations* that potentially expose different concrete interfaces. We therefore needed to model Service Profiles in a high level way that does not prescribe a data model or dictate explicit methods. To do this we created Service Resource Cards (SRCs), based on an existing agile technique called Class Responsibilities/Collaborations first described by Beck and Cunningham (1989). Our SRC models the capability of a service to realise a specific use case. The responsibilities of a service describe at a high level the purpose of a service: what it is for, what it does, and what it can provide to other

components. Collaborations with other services indicate where a service might consume another service to fulfil its own specific use case.



**Figure 2: The Abstract Layers of a SORM**

The Service Profiles and Service Implementations become resources in the model and are interlinked in the same way as other resources. In some cases the functionality of the identified service will be encompassed by existing software systems in the Domain Definition layer, in which case they should be linked together using the appropriate ontological relationship.

*Reference Implementation:* The most concrete layer is an actual reference implementation of the service profiles. Not all services will necessarily be implemented, and some may be wrappers around existing software. The implementations are not intended as definitive enterprise level pieces of code, but as exemplars that validate the service profiles and demonstrate any interoperability (although in open source cases they may also act as an actual software resource). These implementations become the final resources in the Reference Model, and are linked down through the profiles and use cases to the domain definition. This chain of links forms an audit trail that describes exactly why and how the software was conceived. The implementations may also be linked more directly (for example, they may draw on standards, or use software systems that have been described in the domain definition).

Each layer of the reference model is useable in its own right to achieve the use cases from Figure 1:

- *Domain Definition:* This might be used to develop a context for one's own work, to understand how existing work fits together, and to identify standards and locate experts.

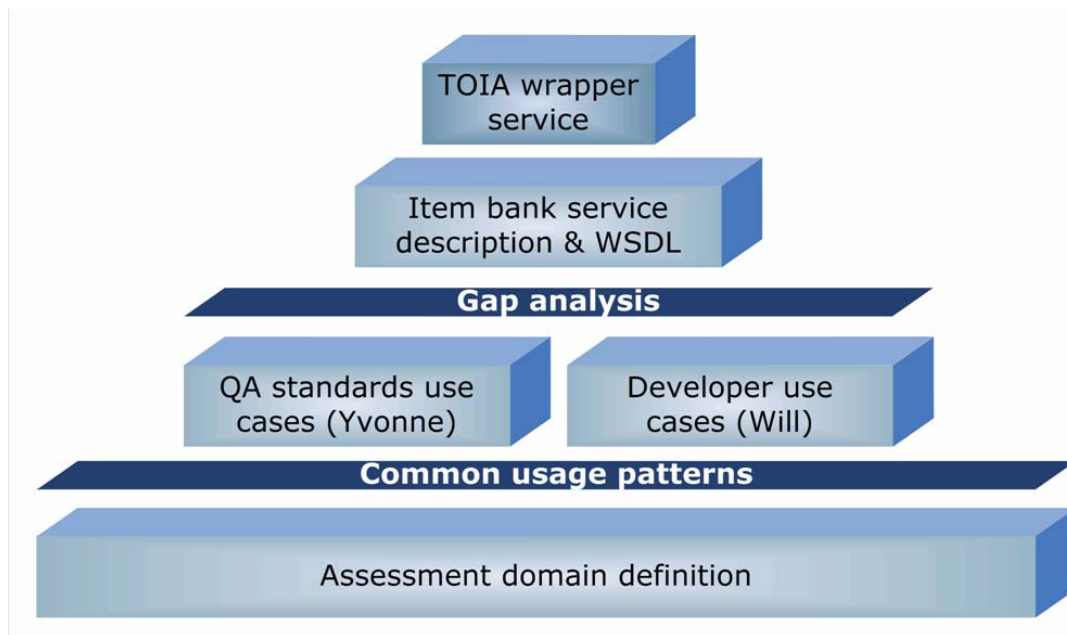
- *Use Cases:* These can be used to help understand usage patterns within the domain and to help developers create new Service Profiles and thus Services.
- *Service Profiles:* Developers that wish to build new services that work within the domain framework will need to use the service profiles to ensure interoperability. They might also wish to create alternative versions of existing services, either to improve on the existing implementations, or for commercial reasons.
- *Reference Implementation:* Finally the actual software implementations are available to those developers that wish to build on some, or all, of the developed services.

We can demonstrate how the reference model supports service discovery and evolution from the perspective of one of our actors, instantiated as a persona, Will, (the domain developer) and follow his activities revealed in the Community Reference Model as he enacts his scenario.

Will's goal is to create and publish new services. He will use the domain definition layer to understand the scope of the domain and follow links into the use case layer to locate where his own expertise lies in the context of use cases and scenarios.

He will use the gap analysis to identify where competition for service delivery is high (many links from use cases to service implementations) and also where there are opportunities for him to create innovative new services (no or few links to service implementations).

For the opportunities he has discovered, Will is able to view service profiles where they have been specified in the service profile layer. Finally, he can follow links from the service profile to a reference implementation for his new service in the implementation layer which shows how it should interoperate with other services in the framework. Will is able to follow the chain of links back to the domain layer to check what domain and technical standards support the service profiles he is interested in. He may also follow the links back to the service profile layer to locate some existing services that he can re-use in the architecture of his new service.



**Figure 3 : Layers of a Community Reference Model from the Assessment Domain**

Figure 3 shows a possible form for our instantiated SORM with our example personas (Yvonne: the Quality Assurance manager interested in standards support, and Will: the eFramework developer). From these uses cases a gap analysis will show which core services need to be profiled, in this case an Item Bank service to support Will’s development. Finally there are reference implementations of these services. In this case there is one, supporting the item bank service and providing a wrapper around TOIA, an existing item bank system.

The full version of the FREMA Community Reference Model is much more complex than this simple diagram is able to convey, with many use cases that map onto many services. It is likely that the FREMA Community Reference Model will both create new service implementations and wrap existing systems (sometimes to reveal more than one service interface).

#### *Structuring the Reference Model*

Since the Community Reference Model is designed to be a community resource it is important that it is accessible to all its users and reveals itself at many levels to them. Because of this required flexibility it is impossible to create a static representation of the resources, and instead we have opted for an ontologically modelled set of resources that are combined dynamically at the time of viewing, allowing different users to see the full domain, from base definitions to final service implementations, from a variety of views.

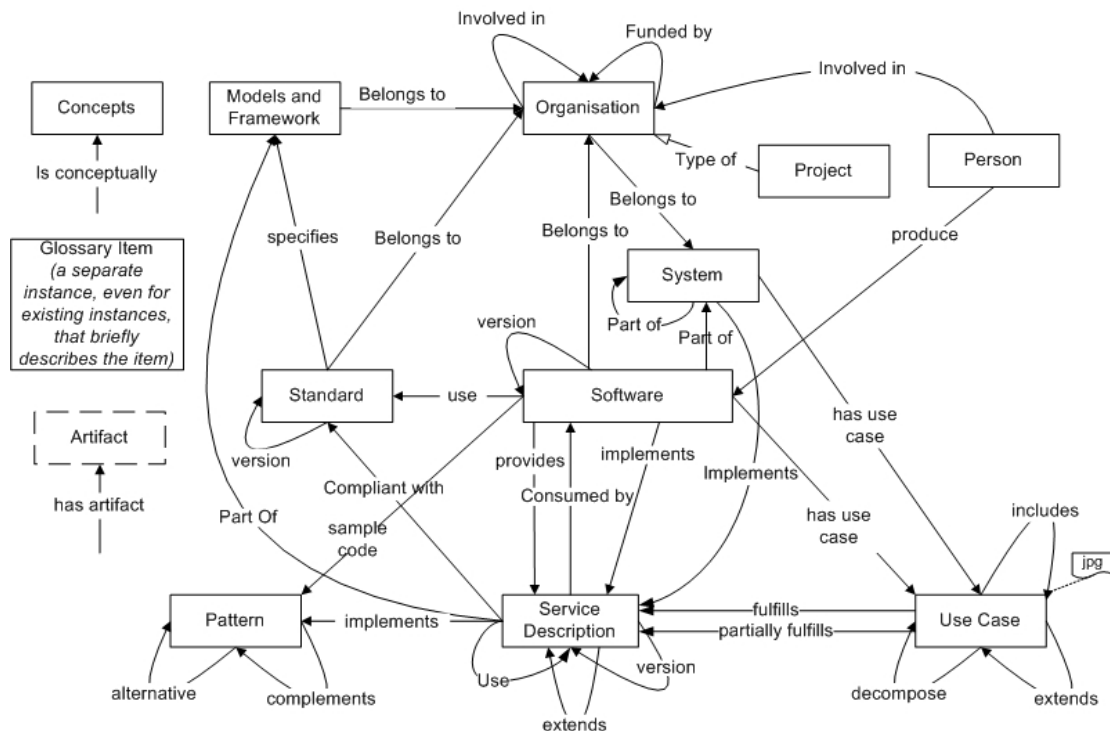
An ontology is simply the collection of classes and relations that are permissible for any given domain (it is called an ontology since it restricts and defines which parts of the world may be understood by entities conforming to it). The advantage of ontological modelling over database schemas is that it enforces a finely grained, and thus flexible and extensible, set of relationships.

It also means that the resources in the Reference Model could be described on the Semantic Web, which would enable interoperability between different Reference Models and reasoning about the described resources.

Figure 4 shows a graphical representation of the ontology that we have used to model our SORM. For simplicity we have not shown the attributes of each class, but have instead concentrated on the relationships between classes.

While the aim of this rich semantic storage layer is to enable users to come into the reference model from many different perspectives, there is a problem in that some users will not know where to find the resources that they are interested in within the model. To this end our ontology also includes a Concept class. Every instance in the reference model can have a conceptual relationship to one or more concept instances. We can then use graphical concept maps to help users orientate themselves and find resources. By investigating various alternative concept maps we hope that we have provided a non-expert means of navigation. We are thus using several kinds of information structure to encode and present the Reference Model.

This ontology is instantiated into a semantic network of resource instances and specific relationships. This is analogous with a *Topic Map* (although topic maps are normally presented visually, and our semantic graph is instead used to generate interlinked web pages). Within the ontology we also model concepts. The semantic graph of these concepts is a *Concept Map* which we do reveal graphically (concept maps can be considered a simple form of topic map that are intended specifically for human viewing and clarity).



**Figure 4: The Reference Model Ontology**

## Navigating the Reference Model

For the FREMA Assessment Reference Model we wanted to use a structure for the domain that could be used by human users of the model to orientate themselves and navigate around the resources. While the underlying resource types are modelled using an ontology, we did not want to expose users to this complexity and we also wanted to avoid the rigidity of a taxonomy. So we chose to create concept maps that described the domain in familiar terms, but which were not explicitly typed or restricted. Every resource in the reference model is associated with at least one concept. Users of the reference model can explore the maps and click through the concepts to the resources that are associated below.

The FREMA concept maps evolved over a period of several months through a series of consultation exercises. We visited a number of community events within the UK and interviewed a number of practitioners with the aim of extracting common terms and perspectives. These informed an initial, informal set of terms and relationships, which we then took back to the community for validation.

Our initial efforts at creating an overview map were a little too complex to be universally understood. We therefore broke down this map over several workshops in an effort to extract a simplified view of the domain. The result was a map of resource types that are considered important within the assessment domain, and a map of the common processes. We refer to the

resource types version as the Noun Map (Figure 5) and the processes version as the Verb Map (Figure 6).

The Noun Map draws heavily from the Ultimate Assessment Engine in that it contains stakeholders and roles (Sclater and Howie, 2003), however because it does not show workflow it does not connect these, or associate them with the types of resources they manipulate. The Noun Map is intended to allow users who deal with specific types of resources to find those resources in the map and discover what other resource types might be relevant.

The verb map shows what people do, but it does not group these activities according to any stakeholders, or relate them to any notion of resource types. There is an implicit clockwise order that follows a common view of how assessments are constructed and executed. The Verb Map is intended to allow people who are interested in a particular activity to find that process, and thus the resources underneath, and also find what other processes are related.

## **Semantic Wiki**

The World Wide Web is the most popular hypertext system, yet it suffers a number of problems when evaluated alongside other hypertext systems. In particular, it has a very clear separation of author and reader, which means that web users cannot change the pages they are viewing. Creating web pages requires specialist skills, and collaborative authoring of a Web site is difficult. One general solution is a WikiWikiWeb (Wiki for short), a type of Web server (or application running on a traditional Web server) that allows any reader of its pages to alter those pages, or create new ones, by using simple web forms (Leuf and Cunningham, 2001). Crucially this allows non-specialist users to contribute to the hypertext.

*Semantic Wikis* (Völkel, *et al.* 2001) are an attempt to use the Wiki concept to make semantics accessible to ordinary users in the same way that ordinary Wikis make hypertext accessible. In Semantic Wikis users are able to type pages and links, forming a semantic network that can be queried. Semantic Wikis make semantics accessible because they are inherently freeform in nature and are non-restrictive, allowing the creation of *semantics-on-demand*, without a complex ontological design process beforehand.

Rather than construct our own Semantic Wiki system, we wanted to exploit an existing system that had typed links, nodes, and first-class types. We looked at a number of existing Semantic Wikis, including IkeWiki (Schaffert *et al.*, 2006). Kaukola (Kiesel, 2006), WikiSar (Aumueller and Auer, 2005) and Semantic MediaWiki (SMW) (Völkel, *et al.*, 2001).

In the end we chose Semantic Media Wiki (SMW) as it is relatively mature (as it is based on MediaWiki), has a large user base, offers a number of Wiki features (such as image and user management) and fits our key criteria.

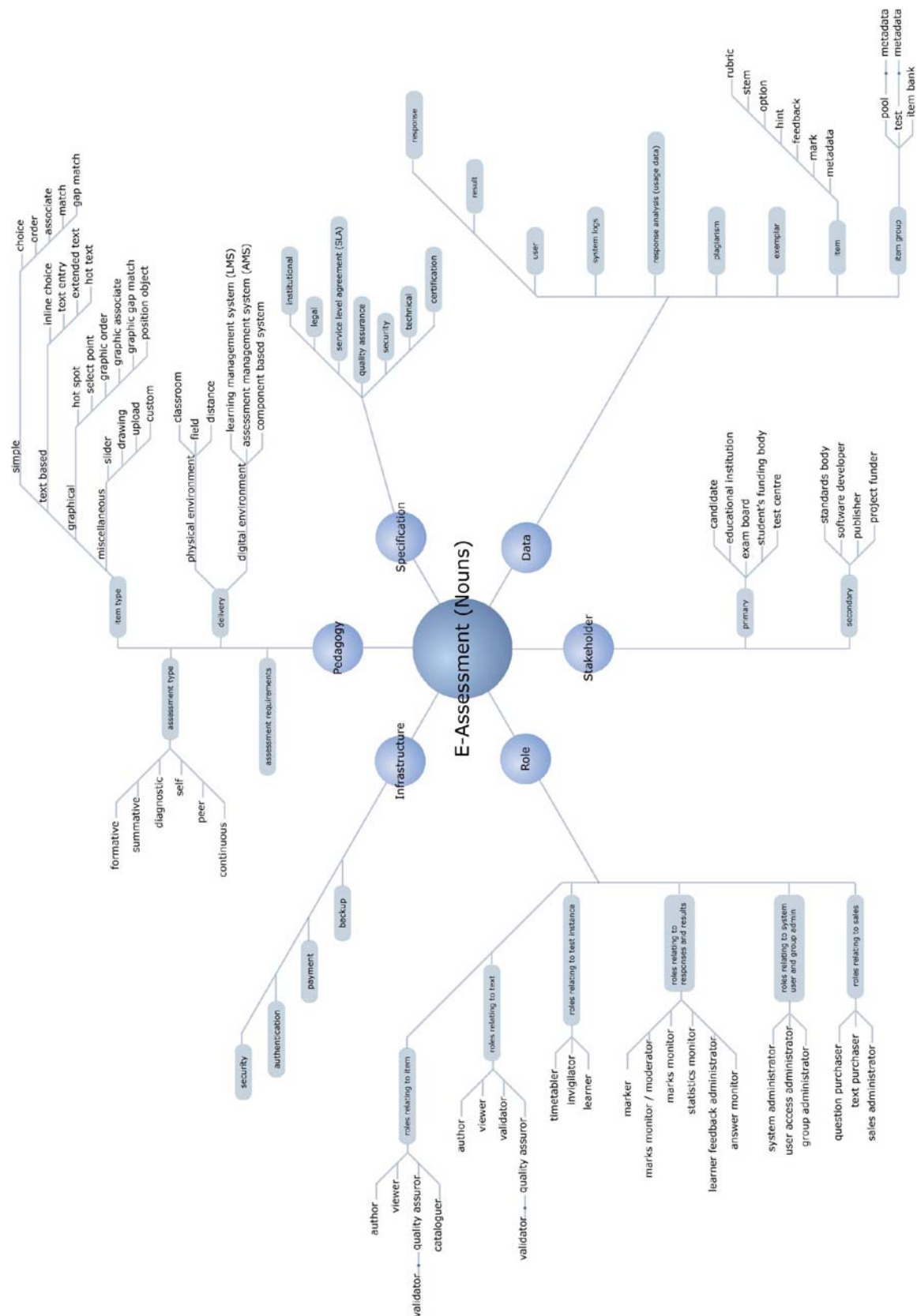
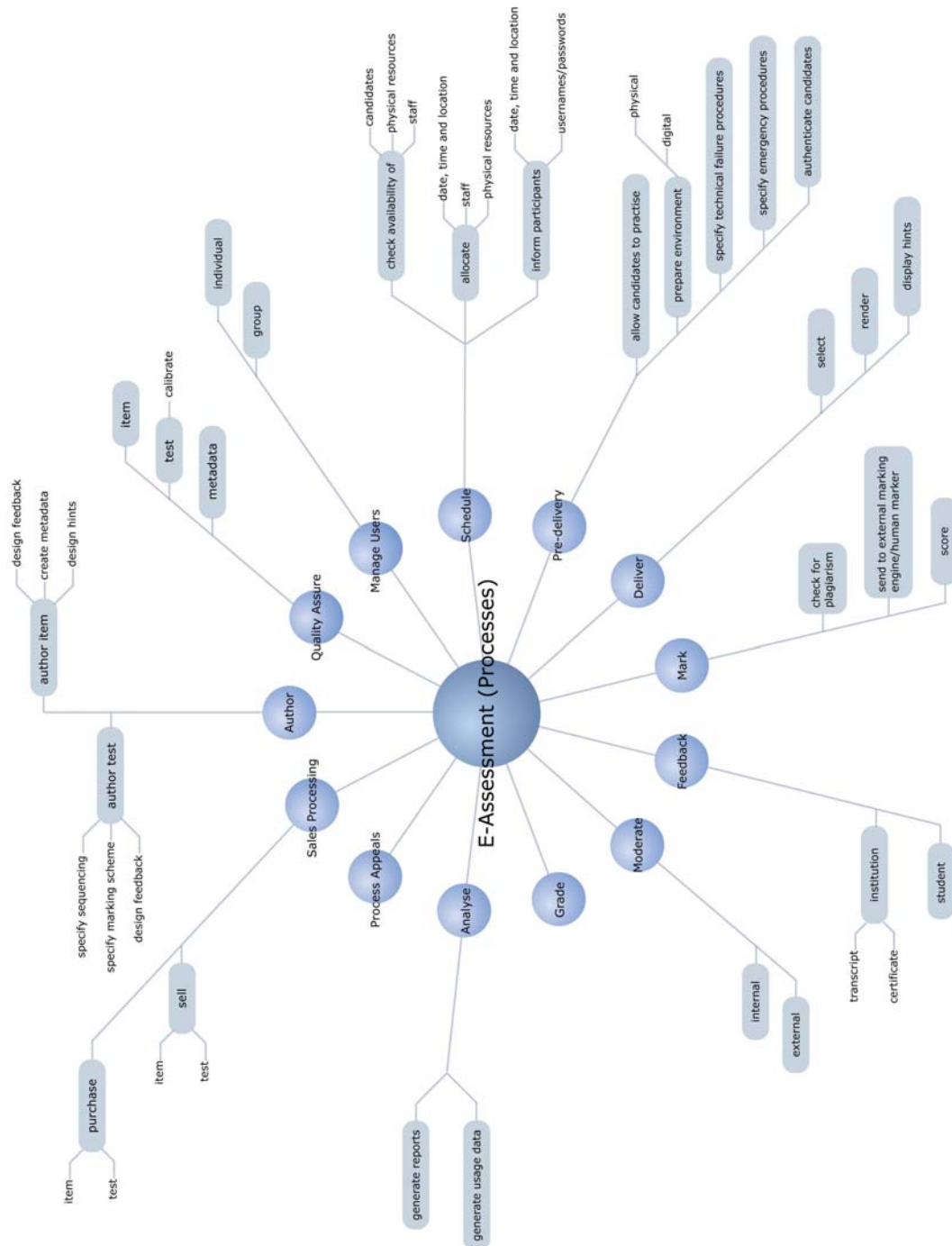


Figure 5: FREMA Noun Concept Map





**Figure 6: Verb Map**

Converting the FREMA knowledgebase into a Semantic Wiki was not the trivial process that we hoped for. However, by sacrificing some of the functionality of the original site, and writing limited SMW extensions, it was possible to replicate most of the original website while gaining all the advantages of using a Wiki: open editing, administration, discussion, file management, etc. Figure 7 shows a list of resources in the FREMA SMW and Figure 8 shows a resource page.

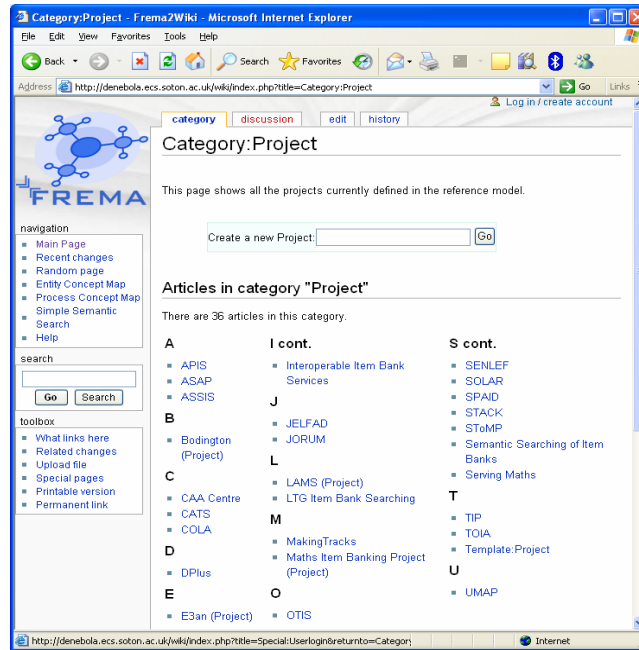


Figure 7 FREMA Semantic Wiki List of Organisations

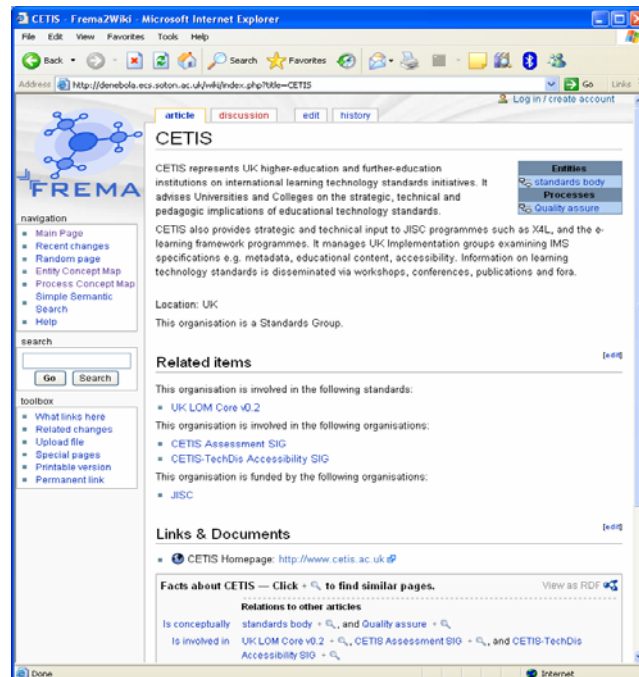


Figure 8 FREMA Semantic Wiki Resource Page

## Conclusion

In any complex domain where disparate people are required to work together to achieve a common aim it can be helpful to provide a mechanism for overview. Once people delve into more specific areas then this becomes a need to support navigation and orientation. We believe that while a complex domain itself may be best modeled with an ontology or a topic map, the

overview is best presented as a concept map. In this paper we have presented two complementary concept maps of the assessment domain that we have used with the FREMA reference model for Assessment. On the FREMA website we use these concepts maps to help users discover resources, orientate themselves within the domain, and discover new areas that might also be of interest.

Moving from a knowledgebase-driven web site to a Semantic Wiki means a change in thinking. One must release control of the structuring ontology, and place one's faith in the wisdom of the user community. But it is also liberating, and the potential advantages are many: a familiar editing paradigm, co-ownership of content, and evolution rather than stagnation of structures and terms.

## References

- Aumueeller, D., and Auer, S. (2005). Towards a semantic wiki experience – desktop integration and interactivity, in WikSAR, *Proc. of 1st Workshop on The Semantic Desktop*, Galway, Ireland.
- Beck, K. and Cunningham, W. (1989). A laboratory for teaching object oriented thinking, *ACM SIGPLAN Notices*, 24(10), 1-6.
- Bloom, B.S. (1956). *Taxonomy of Educational Objectives*, Longman.
- Bull, J., and McKenna, C. (2004). *Blueprint for Computer Assisted Assessment*, Routledge Falmer.
- Chang, W.-C., Hsu, H.-H., Smith, T., and Wang, C.-C. (2004). Enhancing SCORM metadata for assessment authoring in e-Learning, *Journal of Computer Assisted Learning*, 20, 305.
- Cooper, A. and Reimann, R. (2003). *About Face 2.0: The Essentials of Interaction Design*, John Wiley & Sons.
- Conole, G. and Warburton, B. (2005). A review of computer-assisted assessment, *ALT-J Research in Learning Technology*, 13, 17-31.
- Kassam, S. (2004). A summary of the assessment domain, in *Proceedings JISC/CETIS conference*, Oxford, UK.
- Kiesel, M. (2006). Kaukolu - Hub of the semantic corporate intranet, *Workshop: From Wiki to Semantics*, ESWC.
- Leuf, B. and Cunningham, W. (2001). *The Wiki way: quick collaboration on the Web*, Addison-Wesley.
- Olivier, B. (2005). *The eFramework: an Overview*, JISC.
- Olivier, B., Roberts, T., and Blinco, K. (2005). The e-Framework for Education and Research: An Overview, retrieved from [www.e-framework.org](http://www.e-framework.org), July 2005.
- Reich, S., Will, U. K., Nuernberg, P. J., Davis, H. C., Groenbaek, K., Anderson, K. M., Millard, D. E. and Haake, J. M. (2000). Addressing Interoperability in Open Hypermedia: the Design of the Open Hypermedia Protocol, *New Review of Hypermedia and Multimedia*, 5, 207-248.
- Schaffert, S., Bischof, D., Buerger, T., Gruber, A., Hilzensauer, W., and Schaffert, S. (2006). Learning with semantic wikis, *SemWiki 06*, Budva, Montenegro.
- Sclater, N. and Howie, K. (2003). User requirements of the “ultimate” online assessment engine, *Computers & Education*, 40, 285–306

Smythe, C., Evdemon, J., Sim, S., and Thorne, S. (2004). Basic architectural principles for learning technology systems, IMS Global Learning Consortium. .

Wilson, S., Blinco, K., and Rehak, D. (2004). *An e-Learning Framework: A Summary*, JISC.

Wilson, S., Blinco, K., and Rehak, D. (2004b). *Service-Oriented Frameworks: Modelling the infrastructure for the next generation of e-Learning Systems*, JISC.

Völkel, M., Krötzsch, M., Vrandečić, D., Haller, H., and Studer, R. (2006). Semantic Wikipedia, in *Proceedings of WWW 06*, Edinburgh, 585-595.



# **DELIVERY OF QTIV2 QUESTION TYPES**

**Gary Wills<sup>‡</sup>, Hugh Davis<sup>‡</sup>, Lester Gilbert<sup>‡</sup>,  
Jonathon Hare<sup>‡</sup>, Yvonne Howard<sup>‡</sup>, Steve Jeyes<sup>†</sup>,  
David Millard<sup>‡</sup>, and Robert Sherratt<sup>†</sup>**





# Delivery of QTIv2 Question Types

Gary Wills<sup>‡</sup>, Hugh Davis<sup>‡</sup>, Lester Gilbert<sup>‡</sup>, Jonathon Hare<sup>‡</sup>, Yvonne Howard<sup>‡</sup>, Steve Jeyes<sup>†</sup>, David Millard<sup>‡</sup>, and Robert Sherratt<sup>†</sup>

<sup>‡</sup>Learning Societies Lab, University of Southampton, UK.

<sup>†</sup>e-Services Integration, University of Hull, UK

## Abstract

The QTI standard identifies sixteen different question types which may be used in on-line assessment. While some partial implementations exist, the R2Q2 project has developed a complete solution that renders and responds to all sixteen question types as specified. In addition, care has been taken in the R2Q2 project to ensure that the solution produced will allow for future changes in the specification. The paper summarises the rationale of Web services and a Service Oriented Architecture, and then demonstrates how the R2Q2 project integrates into JISC's e-Framework, and the reference model for assessment (FREMA<sup>1</sup>).

The design of R2Q2 is described, the focus being on lessons learnt. We describe the architecture and the rationale of the internal Web services and explain the approach taken in implementing the QTI specification, showing how the design allows for future tags to be added with the minimal of programming effort. A major objective of the design was to solve the problem of having to undertake a major redesign and reimplementation as a result of minor modifications to the specification.

In the 2006 Capital Programme from JISC, three new projects were commissioned in the area of Assessment: one for authoring of items, one for item banking, and one for a complete test engine as described in the QTI specification. The R2Q2 Web service is at the heart of all three projects and this paper will describe how the R2Q2 Web service will be used.

## Introduction

Formative assessment aims to provide appropriate feedback to learners, helping them gauge more accurately their understanding of the material set. It is also used as a learning activity in its own right to form understanding or knowledge. It is something lecturers/teachers would love to do more of but do not have the time to develop, set, and then mark as often as they would like. A formative e-assessment system allows lecturers/teachers to develop and

---

<sup>1</sup> Framework Reference Model for Assessment <http://www.frema.ecs.soton.ac.uk/>

set the work once, allows the learner to take the formative test at a time and place of their convenience, possibly as often as they like, obtain meaningful feedback, and see how well they are progressing in their understanding of the material. McAlpine [11] also suggests that formative assessment can be used by learners to *“highlight areas of further study and hence improve future performance”*. Steve Draper [12] distinguishes different types of feedback, highlighting the issue that although a system may provide feedback, its level and quality is still down to the author.

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. To address this there has been a trend 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. This supports the notion that any of the components could be ‘swapped’ for a better version when it becomes available.

One of the more popular standards that has emerged is Question and Test Interoperability (QTI) developed by the IMS Consortium<sup>2</sup>. 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 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, with no definitive reference implementation [8,9].

This paper presents the Web service R2Q2 and the Test delivery engine ASDEL. R2Q2 is a JISC funded project that brings the SOA approach and QTI standard together to develop a set of Web Services that will render and respond to questions written to the QTI standard. The paper will also report on the progress being made on the ASDEL project, again funded by JISC to develop a QTIv2 compliant test delivery engine.

---

<sup>2</sup> IMS QTI homepage: <http://www.imsglobal.org/question/>

## Service Oriented Architectures

Service-Oriented Architectures (SOAs) enable large complex systems to be mutualised, that is composed of independent software components that operate through well-defined interfaces. 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: the ability to dynamically couple services, interoperability of services due to clearly defined standards, and as a result the ability to avoid 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.

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, 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. JISC call this initiative simply the 'e- Framework'.

The e-Framework is based on a service-oriented factoring of a set of distributed core services [17], where flexible granular functional components expose service behaviours accessible to other applications via loosely coupled standards-based interfaces. The technology used is Web Services and the intention is to extend the SOA programming model into a vast networking platform that allows the publication, deployment, and discovery of service applications on the scale of the Internet.

For the assessment domain, the reference model is FREMA (Framework Reference Model for Assessment)<sup>3</sup>. 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].

## Question and Test Interoperability

The IMS QTI Specification is a standard for representing questions and tests with a binding to the eXtended Markup Language (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

---

<sup>3</sup> FREMA homepage: <http://www.frema.ecs.soton.ac.uk/>

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.

---

```
<?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>
      
    </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)**

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) Hotspot            |
| 3) Order          | 4) Select point       |
| 5) Associate      | 6) Graphic            |
| 7) Match          | 8) Graphic Order      |
| 9) Inline Choice  | 10) Graphic Associate |
| 11) Text Entry    | 12) Graphic Gap Match |
| 13) Extended Text | 14) Position object   |
| 15) Hot Text      | 16) Slider            |

The list of different question types can be combined with templated question or adaptive response profiles, providing an author with numerous alternative

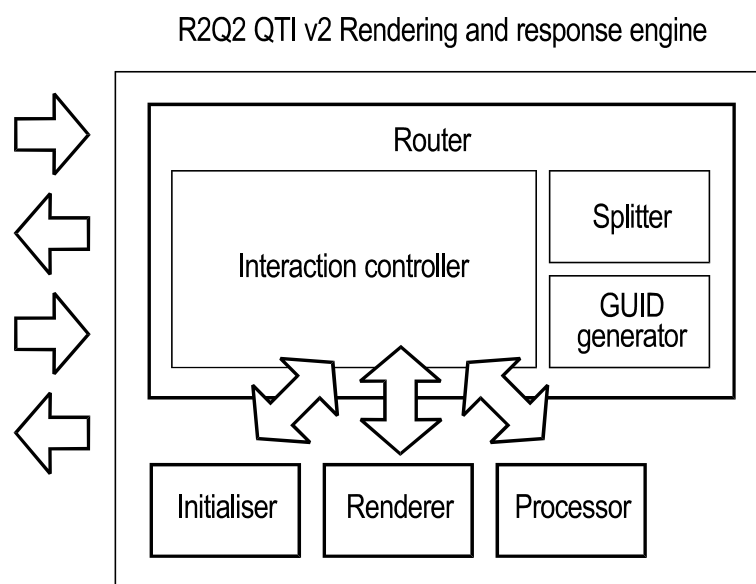
methods for writing questions appropriate to the needs of the students. Templated questions include variables in their item bodies that are instantiated when a question is rendered (for example, inserting different values into the text of maths problems). Adaptive questions have a branching structure, 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.

## R2Q2 Design

The R2Q2 service allows a student to view a question, answer a question, and view the feedback. The R2Q2 engine (see Figure 2) 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, maintaining 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.



**Figure 2 The R2Q2 Architecture**

## Integration into a Portal framework.

Figure 2 shows the core services where R2Q2 is used as a stand alone service. To ensure wide-spread take up of the Web service, R2Q2 is also designed to be dropped into applications such as a VLE, portal framework, and test engine authoring tool, amongst other applications, to achieve the aim of migrating the community to this new standard. To this end the project Web site provides documentation for installation, and a single install process.

When integrating Web services with VLEs and portal frameworks, we have found that you cannot just call a service, but code needs to be written to manage calls to and information from the Web services. The generic name for such a piece of code is an adaptor (see the EFSCE project<sup>4</sup>).

The R2Q2 project provides a demonstrator in the form of a Web client that uses traditional XHTML and JAVA servlets to display the questions. There are key differences to be considered between a portlet implementation of R2Q2 and a more traditional simple servlet implementation. The java PortletRequest object involves a protocol which is different from that of a HTTPServletRequest object. The main difference is that the portlet requests contain additional information regarding the portlet window within the portal. As a result, the way the request is handled will be different, for example within the R2Q2 demo it is no longer possible to use the ServletFileUpload class as a file upload handler for the request.

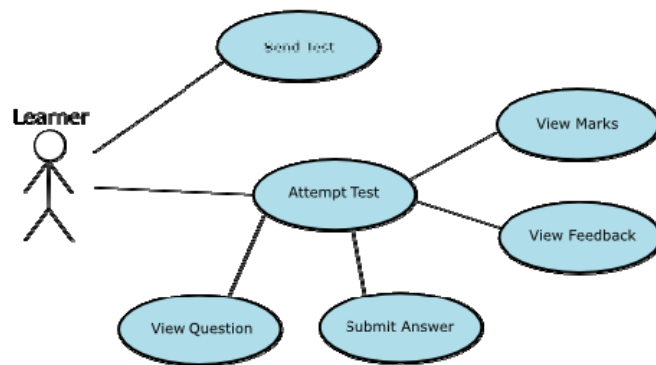
There are a number of open source portal frameworks that are currently being used. They are all similar in that they are Java-based and use a Model View Controller (MVC) architecture. The MVC architecture separates the presentation code from the business logic code and is implemented using Struts for web applications. Struts provide a mechanism by which the flow of information is directed to the correct portlet. The way this is implemented means that the system can scale quite easily. Struts model the various functions of the portlet as 'actions'. When an action URL is sent, a controller redirects the portlet to the correct JSP page which connects to the Web service.

## ASDEL

R2Q2 successfully implemented a rendering and response engine for a single question (also termed an item), for which there are sixteen types described in the specification and implemented in R2Q2. While this is useful, it does not implement the whole of the QTI specification regarding the test process. The specification details how a test is to be presented to candidates, the order of the questions, the time allowed, etc. The typical use-case from the point of view of a learner candidate of the test process is illustrated in Figure 3.

---

<sup>4</sup> EFSCE project Web page <http://www.efsce.ecs.soton.ac.uk/overview>



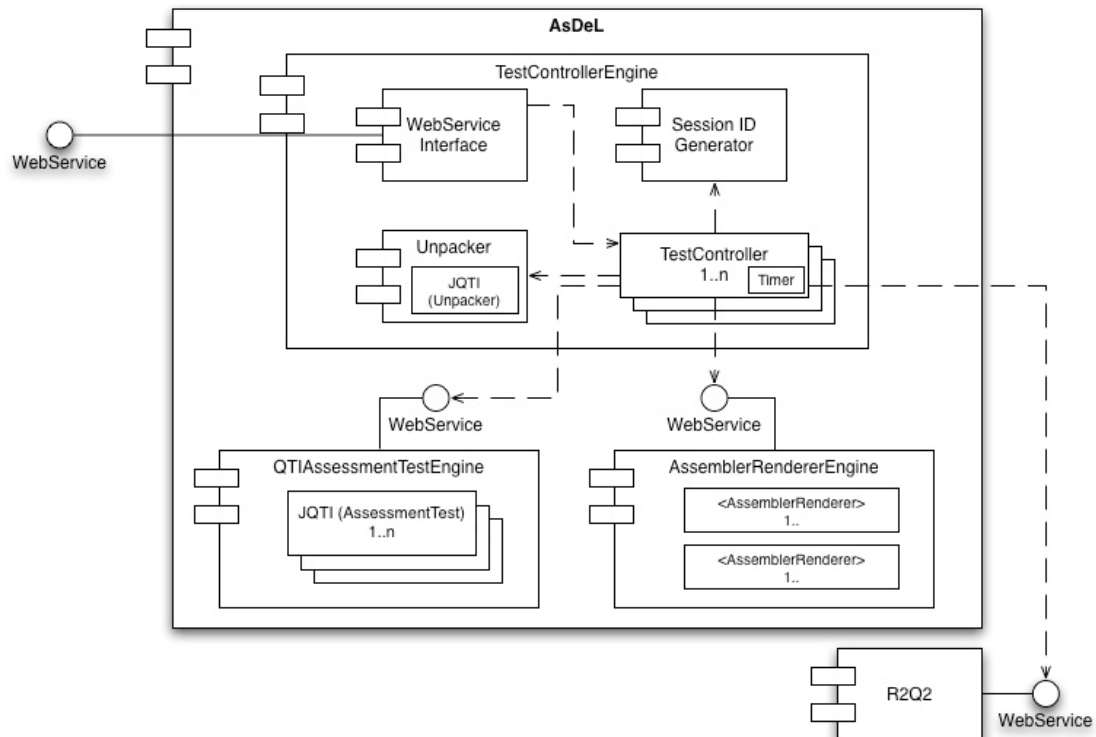
**Figure 3: Use case of ASDEL from the user perspective**

In the ASDEL project we aim to build an assessment delivery engine to the IMS Question and Test Interoperability version 2.1 specifications that can be deployed as a stand-alone web application or as part of a Service Oriented Architecture enabled Virtual Learning Environment or portal framework. The engine will provide for:

- Delivery of an assessment consisting of an assembly of QTI items, with the possibility that the assessment is adaptive and the ordering of questions can depend on previous responses,
- Scheduling of assessments against users and groups,
- Rendering of tests and items using a web interface,
- Marking and feedback,
- A web service API for retrieving assessment results.

Like R2Q2, the ASDEL project will use a Service Oriented Architecture (SOA). The design of the ASDEL system specifies that the major components will be created as internal Web services.

Phase 1 is the technical development of the engine in accordance with the IMS QTIv2.1 specification and in accordance with the JISC e-Framework approach of using web services in a Service Oriented Architecture (see Figure 4). The engine will take in a test as an IMS Content Package or by reference to the test XMLfile. The engine will unpack the content package and assemble the items into a directory on a local file system. The engine will import any additional material (images, videos, etc) required by the test, and it will then process the XML and deliver the test as scheduled to the candidate via a Web interface. Feedback will be given to the candidate and the marks processed in accordance with the schema sent to the engine. The results can be retrieved through the engine API. The engine will also have the additional features of being able to persist partially completed tests for future completion, and the ability to record candidate responses (in addition to results) for later review.



**Figure 4. Architecture for the Assessment Delivery system.**

The core components of the ASDEL system will be built around a Java library, which has been termed JQTI. The JQTI library will enable valid QTI assessment XML documents to be interpreted and executed. The library will also provide auxiliary services like the handling of QTI content packages and the provision of valid QTI conformance profiles and reports.

The AssemblerRenderingEngine part of the system is responsible for the assembly and rendering of output (i.e. questions and associated rubric). Initially, only an XHTML renderer will be developed; however, the design of the engine will enable different renderers to be plugged in.

Figure 5 illustrates the typical sequence of events when a user is interacting with the ASDEL system through a particular portal or VLE. Figure 6 shows the typical initialisation stages that the system goes through when a test package is presented, and Figure 7 demonstrates the typical collaborations between system parts when the a learner is undertaking a test.



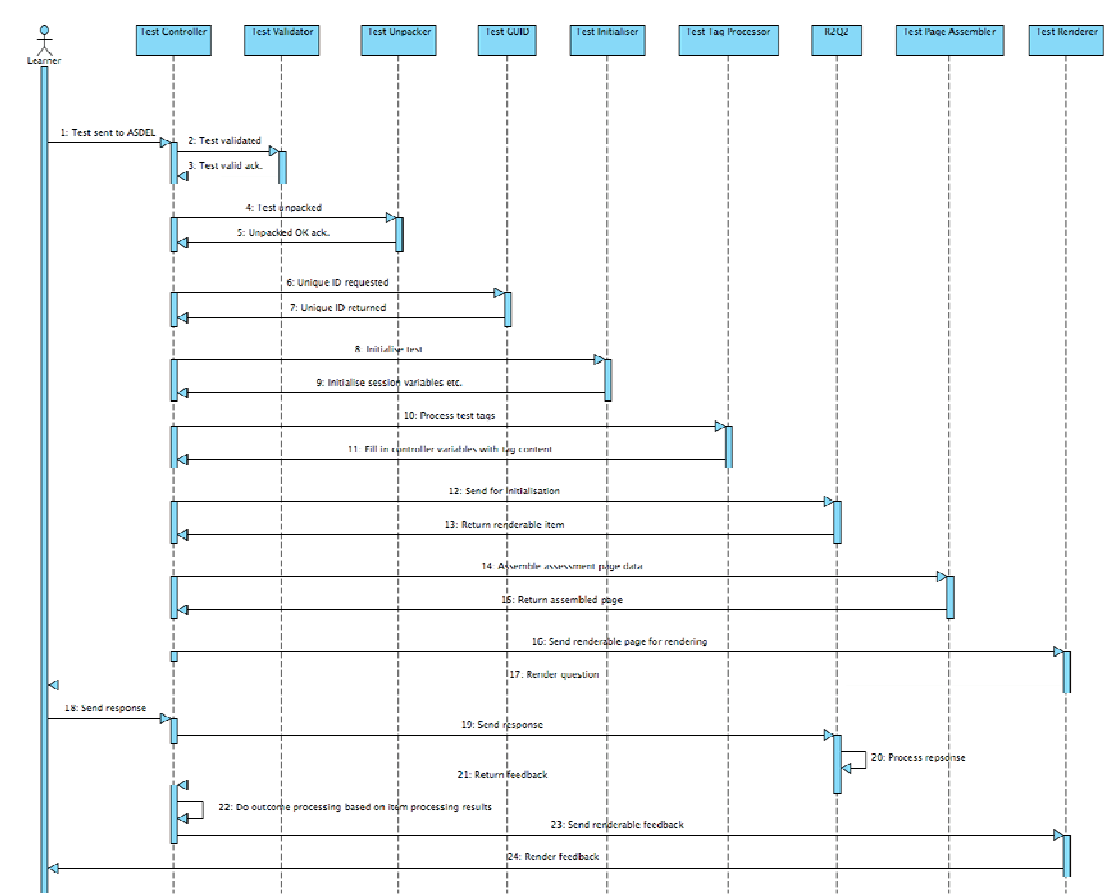


Figure 5: Typical sequence of events within the ASDEL system

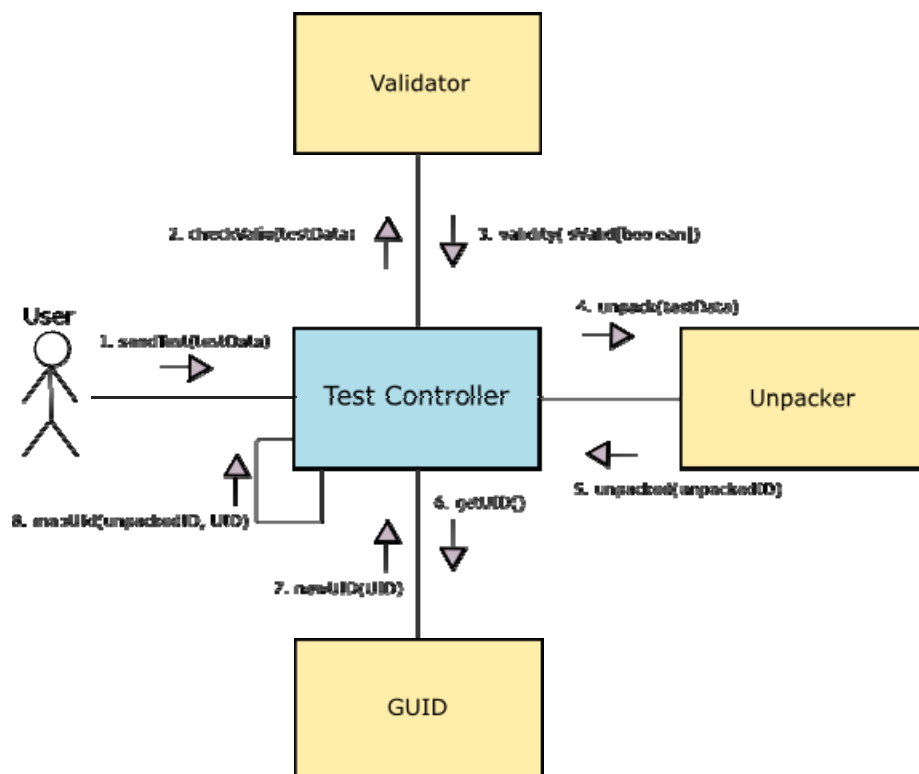


Figure 6: Collaborations between components during initialisation of a test

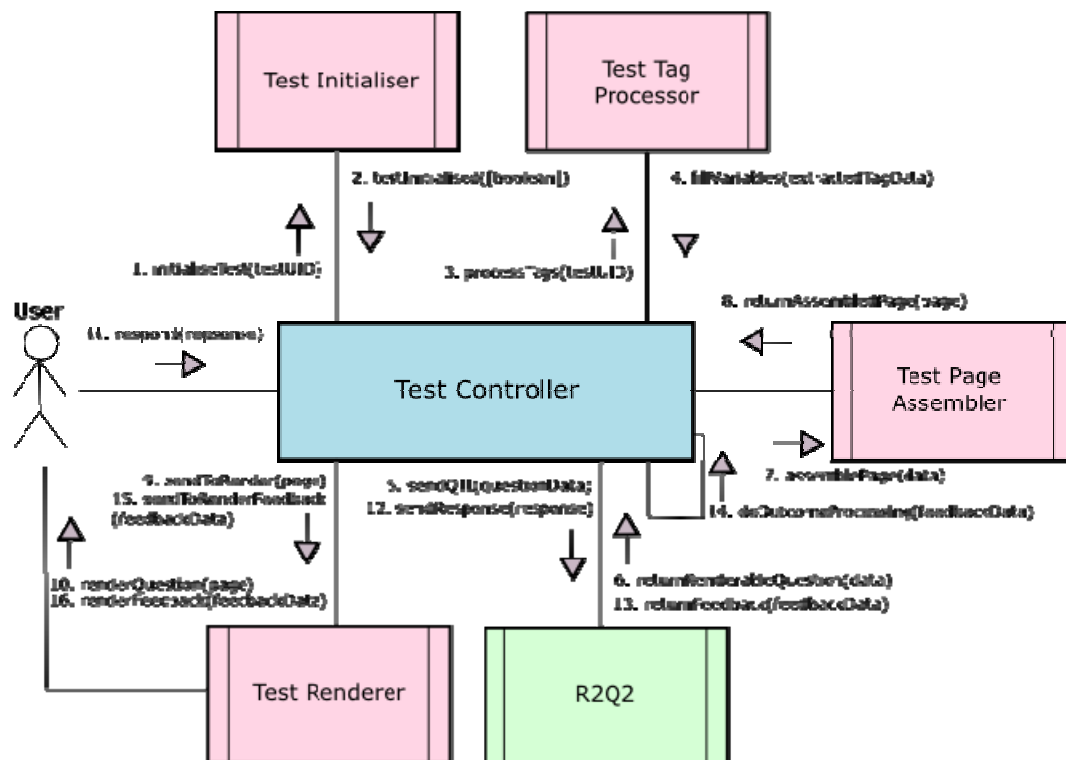


Figure 7: Collaborations between components as a test is undertaken

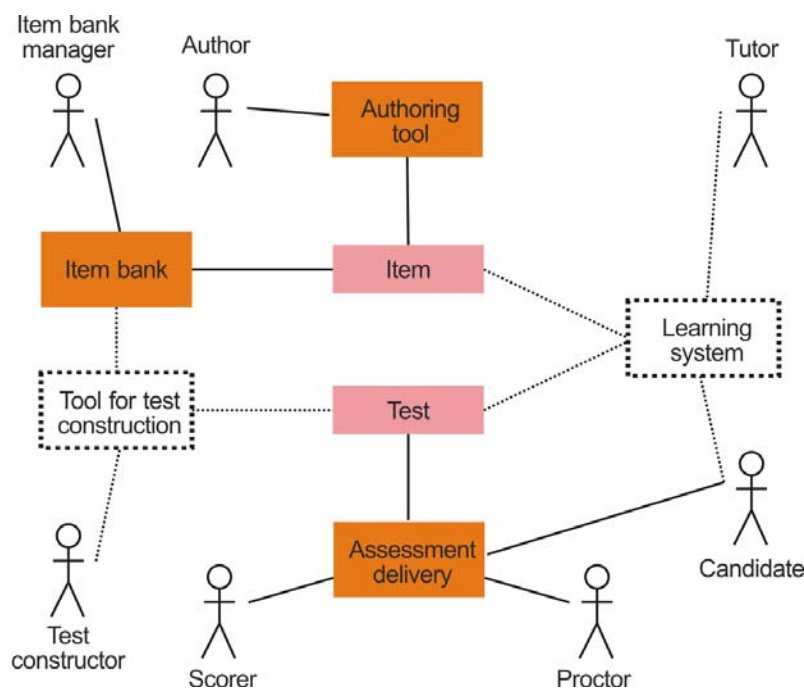


Figure 8. Phase Two: Integration of the ASDEL, AQuRate Item Authoring (Kingston) and MiniBix, Item Banking (Cambridge).

In the second phase, the project will integrate with the other projects in the JISC Capital Programme call on item banking (Cambridge: Minibix) and item authoring (Kingston: AQuRate) to provide a demonstrator, and will contribute to its evaluation and the evaluation of the project.

**Figure 8** shows a modified diagram of the Use Case from the QTI v2 specification, demonstrating how the different tools and system in this call relate together. It clearly shows the boundaries between the delivery system, authoring tool, and item banking. A general scenario would be:

1. A lecturer/tutor will write questions (items). The authoring tool will provide a user interface appropriate to the end user, and format and store the items using the QTI v2 standard. By using QTIv2 these items may be exchanged with other compliant systems not developed by the same developer.
2. Users can select items from the item bank and place the items in a pool ready for constructing into a test. The test construction system, like the item authoring tool, will use an appropriate user interface and behind the scenes output the test in a QTI v2 or IMS CP compliant format.
3. By having the test and item adhere to the QTIv2 specifications, the deployment of items, item banks, and tests from diverse sources can be delivered through the test delivery system to candidates via a learning environment or directly via their internet browser.
4. The candidate can now take the test, and have the results reported in a consistent manner.

The integration in this workpackage may be best achieved by using a portal framework to integrate the different projects.

## **Changes to R2Q2**

During the design and implementation of ASDEL a number of issues have been identified in R2Q2 that will need to be fixed before the implementation is complete. Firstly, the default R2Q2 renderer renders full xhtml pages rather than rendering fragments. ASDEL requires fragments so that it can append various elements of rubric and other textual information about the test before and after the question. In the bigger picture, the output from ASDEL also needs to be in the form of a fragment so that it can be integrated with a VLE or portal framework. The second issue is that R2Q2 will always render the feedback that is included in an item. The problem is that the QTI assessment specification allows the delivery engine to control whether or not the feedback from an individual item should be delivered.

## **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.

R2Q2<sup>5</sup> is a definitive response and rendering engine for QTIv2 questions. While this only deals with an Item in QTI terms, it is essential to all processing of QTI questions; that is, it forms the core component of all future systems. Due to the design and use of internal Web services, the system could be enhanced if required. So while every effort has been made to ensure this service can be dropped into future systems, if necessary it can be changed to suit any application. The R2Q2 rendering and response engine of QTIv2 questions is expected to help two main stakeholders:

- **Early adopters of QTIv2** have written questions to this specification and need to validate the question. To help them we have provided a Web client to which they can submit questions and see the rendered version.
- **Other e-Framework Projects.** We have provided the core elements of QTIv2 appropriate to a service oriented architecture. Applications in the area of e-assessment, and other aspects of the specification, need to be developed. The R2Q2 project would be an essential element in such future work.

In the ASDEL project we aim to build an assessment delivery engine to the IMS Question and Test Interoperability version 2.1 specifications. Like R2Q2 this will be a Web service based system that can be deployed as a stand-alone web application or as part of a Service Oriented Architecture enabled Virtual Learning Environment or portal framework. The engine will provide for:

- Delivery of an assessment consisting of an assembly of QTI items, with the possibility that the assessment is adaptive and the ordering of questions can depend on previous responses,
- Scheduling of assessments against users and groups,
- Rendering of tests and items using a web interface,
- Marking and feedback,
- A web service API for retrieving assessment results.

---

<sup>5</sup> <http://www.r2q2.ecs.soton.ac.uk/>.

## 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](http://www.e-framework.org), accessed July 2005.
- [6] Sclater, N. and 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.
- [11] McAlpine, M. Principles of Assessment, Bluepaper Number 1, CAA Centre, University of Luton, February 2002.
- [12] Draper, S. W. Feedback, A Technical Memo Department of Psychology, University Of Glasgow, 10 April 2005: <http://www.psy.gla.ac.uk/~steve/feedback.html>.



**COMPUTER ASSISTED TESTING OF  
SPOKEN ENGLISH: A STUDY TO  
EVALUATE THE SFLEP COLLEGE  
ENGLISH ORAL TEST IN CHINA**

**Xin Yu and John Lowe**





# **Computer Assisted Testing of Spoken English: A Study to Evaluate the SFLEP College English Oral Test in China**

Xin Yu and John Lowe  
University of Bath

## **Introduction**

'If you want to encourage oral ability, then test oral ability' (Hughes, 1989:44)

Since its opening up to the outside world in the 1980s and the introduction of economic reforms that have involved engagement with the global economy and wider community, the Chinese government has become determined to promote the teaching and learning of English as a foreign language among its citizens. In particular, it has mandated the study of English for all college and university students and has made the passing of the College English Test (CET) at Band 4 level a requirement for obtaining a degree. With some ten million candidates annually (and rising) CET Band 4 has become the world's largest language test administered nationwide (Jin and Yang, 2006). In a deliberate attempt to harness the backwash effect of examinations on teaching and learning, the Ministry of Education has insisted that all college and university students (generally when in their second year of study) must sit the CET Band 4 written papers that test reading, writing and listening skills in English. Aimed largely, but not exclusively, at those students majoring in English, there is also a higher level, Band 6, CET available.

A problem arises, however, when it comes to the formal testing of spoken English. There is a CET Spoken English Test (CET-SET), in use since 1999, which uses the widely accepted format for such assessments of a face-to-face interview with an examiner, together with a discussion on a given topic with two or three other students taking the test at the same time. This approach is both labour- and time-intensive, however, demanding highly skilled examiners as interlocutors and 'small batch' examining of students in sequence. As a consequence, simply for practical reasons of manageability of the test, CET-SET is only available to those who score higher than 80% in the Band 4 written tests or 75% in the Band 6 tests. Slightly conflicting data are available on the numbers taking this speaking test, with the lowest figure seen being around 40 000 and the highest around 90 000 (Jin and Yang, 2006: 22 & 30; Yang, personal communication, 2006). Whatever the precise figure, these data do indicate that over 99% of those students taking CET Band 4 written papers are not taking a test of spoken English. Even for those who do take the CET-SET, the stakes are not so high, since passing this test is not mandatory for obtaining a degree, unlike the CET written papers. The backwash implications of this are clear: neither among students learning English nor

among teachers teaching it in China's colleges and universities is there an emphasis on the development of spoken English proficiency. The high-stakes nature of CET Band 4 means that reading, writing and listening skills are taken seriously, but speaking skills receive much less attention, if any.

With this situation in mind, and with the recognition that computers are now a common feature of the higher educational environment in China, the Shanghai Foreign Language Education Press (SFLEP) and the University of Science and Technology of China (USTC), in Hefei, have been developing, since 2004, a computer-assisted speaking test, the SFLEP College English Oral Test System. (This produces the hideous acronym, SFLEPCEOTS, which will be substituted by CEOTS for the rest of this paper!) This test system is out of necessity, given current limitations of speech recognition software, something of a half-way house towards a fully computer-based assessment of speaking. It removes the need for a skilled examiner to be present during the conduct of the test but still requires such an examiner for the grading of the students' performances. Students sit at a computer, log into the test system after a security test, and then respond to instructions on the screen. The test itself provides a variety of situations to which students respond in spoken English. Examples and details of the test items will be given during the presentation of this paper but not here. They include, however, responses to text, pictures and video clips, and even a discussion with two or three other students, randomly linked together. The students' responses are recorded and then analysed and graded by examiners when they log into the system later. USTC use of this system over the last two years has shown that over 1500 students can take the test and have their performances graded in two or three days. It is argued, therefore, that CEOTS may present a more efficient system than the traditional face-to-face oral assessment and make regular testing of speaking proficiency on a large scale possible, while meeting the universities' daily teaching needs in terms of its usability.

The University of Science and Technology of China has carried out some evaluation of the testing process, in terms of students' perceptions and also of inter-marker reliability. This paper reports on a proposed joint study by USTC, SFLEP and the University of Bath, that will engage in a more thorough and wide-reaching evaluation of various aspects of this system and the possibility that it may offer an alternative to the current CET-SET that will open up the testing of speaking competence to the majority rather than a tiny minority of college students. This study is in its early phases and this paper will discuss some underlying conceptual issues and outline an evaluation research agenda, but will not report any of the provisional pilot data that have been collected so far. Although a considerable amount of data has been collected by USTC through the use of the test over two years, these data were not collected with a systematic evaluation of key aspects of the system in mind and it is recognised that further systematic data collection of various sorts is required. We would like to recognise at this point the generous support that has been afforded to the two presenters by professors Wu Min and Li Mengtao at USTC, who have been largely responsible for the development of CEOTS; but also to other colleagues at SFLEP and the National College

English Testing Committee who have offered and provided support to the development and implementation of the study.

### **Aims and Objectives of the Study**

The research project that we are developing with our partners aims to evaluate the SFLEP College English Oral Test System with a specific concern over its potential for use as part of the CET programme in Chinese universities. It is recognised, however, that our findings may also have more generic implications for the use of computer-assisted English speaking tests, particularly with regard to the promotion of spoken English in Chinese universities. In order to gain acceptance for the test's large-scale public use we must establish its reliability and consider issues of efficiency and test manageability. If the test is to be acceptable as a replacement for traditional face-to-face oral English testing, however, the central concern is its comparative validity, and it is with issues of validity that this paper will primarily deal. Achievement of the additional aspiration to promote a positive backwash effect on English language teaching and learning by encouraging serious attention to be paid to spoken English depends not directly on the nature of the test itself but on whether it is adopted for high stakes use. This will depend on whether it becomes included with the tests of other language skills that must be passed in order to graduate. While willingness for such inclusion by the authorities – notably the CET board - will certainly depend on establishing the test's reliability and manageability, we argue below that at heart this remains a validity issue, drawing on Messick's concept of 'consequential validity' and the social impact of testing.

Specifically, the research questions that we have initially identified as guiding the evaluation of CEOTS are:

- How do the reliability and validity of the SFLEP College English Oral Test System compare with methods of face-to-face testing?
- Is the system efficient and manageable for use with very large numbers of students?
- What are the perceptions among users – both teachers and students – of the impact on English language teaching and learning of the introduction of this system?

The use of computer assisted tests of speaking proficiency is a relatively new field and, as yet, no large and detailed studies have been carried out to investigate the issues that we have identified, especially in relation to the situation in China. The first of the three questions above raises some broad issues that bring together three distinct fields: assessment, linguistic analysis and human-computer interaction. Before discussing a possible research agenda and methodological approach to address the research questions, it is important to identify concepts and theoretical approaches within these fields that we feel are particularly important.

## Assessment: validity as a central concern

As Bachman and Palmer (1996) point out, the ideal outcome to any assessment regime is to achieve a balance among the essential qualities of validity, reliability, impact, and practicality to meet the requirements of the testing context. These qualities – or variations on them (e.g. Gipps 1994) – might usefully be taken to be the components of an evaluation of the assessment's 'fitness for purpose. Wolf (1998) identifies validity as being widely treated as the most crucial consideration in assessment. We concur and feel that - while recognising the importance of other concerns, particularly in a high-stakes context – validity remains the most significant issue in the context of CEOTS and its use in China. Our case depends, however, on a careful and contextualised interpretation of the concept of validity.

Traditional conceptualizations of test validity derived from psychometric testing (e.g. APA, AERA & NCME, 1966) treated validity in terms of three distinct facets, or evidential areas: construct validity, criterion validity and content validity. But, according to Messick, such a view is inadequate:

*'Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment'(Messick, 1989:13).*

Validity as it is more widely understood today is an argument justifying certain interpretations to be drawn from or actions to be based on test scores; it is not actually the test that is valid, but rather the interpretations, conclusions and actions based on the test scores (Roever & McNamara, 2006). The crucial issues of test validity are 'the interpretability relevance, and utility of scores, the important or value implications of scores as a basis for action, and the functional worth of scores in terms of social consequences of their use' (Messick, 1989:13). Black (1998) highlights the fact that all assessment processes are fundamentally social in character. Therefore test users should consider questions not only of how accurate a measurement is but also questions such as 'how valid are the interpretations made from the test data?' and 'how valid are the tests in terms of the decisions that are to be made?' As Kyriakides (2004) points out, validation should deal with issues concerning the consequences of test use (Kyriakides, 2004).

Messick's notion of consequential validity is central to making our case for the need to consider alternatives to face-to-face testing of spoken English in the Chinese CET context, in order to be able to assess the oral English competence of all students and not just a tiny proportion. It brings the backwash effect of the current CET-SET arrangements within the validity argument that has frequently been used to justify face-to-face testing as the 'most valid' form of assessment of speaking proficiency. This argument is generally based on an understanding of the construct validity of speaking tests that construes them as being more 'authentic' in their representation of the 'real-life' use of spoken language. We shall look at the construct validity of such tests in a moment but it is worth noting O'Loughlin's (2002) point that a

language test need not reflect all aspects of 'real-life' communication (including gendered difference) in order still to be valid.

Tests are valid only for specific purposes (Madaus & Pullin, 1991) and in view of the complexity of the validation process, the suggestion that a test's use or purpose should serve as a guide to validation is accepted (Worthen, Borg & White, 1993; Read & Chapelle, 2001). A difficulty clearly emerges when a test is intended to serve more than one purpose. In such a case it is likely that some form of compromise towards 'optimal validity' in relation to all the intended purposes must be sought. In the case of the CET system in China, two purposes may be identified. On the one hand, the test is intended to certify the individual's competence as a speaker of English; on the other hand the test is intended as part of a strategy to promote more effective teaching and learning of English communication in all forms: reading, writing, listening and speaking. Clearly the two are not entirely separable in that if the speaking test does not validly certify spoken English competence but some other skill, it cannot serve to promote the teaching and learning of that competence. Thus, the argument goes, construct validity is the prime form of validity with which we must be concerned; although in the Chinese context outlined here, some way of dealing with the consequential validity issue of backwash also clearly needs to be sought. An important starting point in attempting to deal with this apparent tension in the purposes and their implications for validity is, therefore, to clarify how 'construct validity' might be understood in relation to the testing of spoken language.

### **Linguistic analysis: communicative competence**

A speaking test can be defined as 'a test in which a person is encouraged to speak, and is then assessed on the basis of that speech' (Underhill, 1987:1). This minimal definition does not contain the point that any act of speaking serves a purpose, that purpose being communication. The 'communicative competence' approach to the teaching of language widely predominates in current practice; the replacement of more traditional, grammar and textual analysis models of language teaching by one which focus on developing communicative competence has been a major recent development in language classrooms in China. In relation, therefore, to a context of learning and teaching English, it seems reasonable that our interpretation of construct validity should be based on communicative competence models of language use and learning. Indeed, Heaton (1988) argues that construct validity assumes the existence of certain learning theories or constructs underlying the acquisition of abilities and skills. The case is even stronger if we adopt Caroline Gipps's (1994) suggestion that in an educational assessment regime 'curriculum fidelity' (where curriculum is to be interpreted broadly and not just in terms of subject content) is a more useful concept than that of construct validity that arose from the psychometrics testing tradition.

This is not the place to present a full account of communicative competence and its meaning for language teaching and learning, but the table below provides a useful summary of key aspects of the approach and a framework

within which we may start to consider the construct validity of any form of speaking test that is intended to serve a communicative competence based curriculum.

**Table 1: The components of a communicative competence model of language**

Grammatical competence	Mastery of the language code
Sociolinguistic competence	Knowledge of appropriate language use
Discourse competence	Knowledge of how to connect utterances in a text so it is both cohesive and coherent
Strategic competence	Mastery of the strategies that speakers use to compensate for breakdowns in communications as well as the strategies they use to enhance the effectiveness of the communications

(Based on Canale and Swain 1980)

In principle at least, any given form of language assessment – whether of speaking or another skill – can be examined in relation to the extent to which it provides the opportunity for the candidate to demonstrate each of the above competences. But we should also be prepared to accept that no one form of assessment will assess all of the components equally well. As with any form of assessment, some sort of sampling of the domain will have to take place. This is revealed after a moment's thought about 'sociolinguistic competence', for example: clearly, even just within the domain of spoken language, the range of 'appropriate' language uses is enormous and way beyond the capacity of any manageable single assessment instrument to do anything more than lightly sample. In our comparisons of face-to-face and computer assisted modes of assessment we shall adopt the principle of asking what it is that each assesses, from the communicative competence model, rather than prioritising any component of that model in advance, thereby fitting an approach to validity that asks what interpretations can be made of performance in the assessment tasks.

Obviously, in the SFLEP College English Oral Test System, using computers instead of the interlocutors of a face-to-face test changes the participants in the speaking course. One of the participants in the interaction is changed to a computer, which may have potential effects on the students' language output. Almost no comparable research has been done between face-to-face and computer-assisted speaking tests, while there is some literature comparing tape-recorded tests with face-to-face speaking tests. The availability of visual as well as auditory stimuli is a key difference between computer based tests and those based on a recorded voice alone.

Kraut et al (1990) suggest that the visual channel is necessary to initiate a conversation in informal communications. As people talk, they are seeking positive understanding, such as acknowledgements, which take the form of gestures such as head nodding (Goodwin, 1981). Modes of body language, such as head gestures and facial expressions are well known to have strong effects on interactions in social situations generally (e.g. Argyle, 1983). It is therefore possible that the visual channel can affect the actual assessment of students' answers in an oral testing situation. For example, gestures of the hand could amplify a spoken explanation to advantage; whereas frowns could predispose the assessor in an unfavourable way towards a student (Seddon and Pedrosa, 1990). This has given support to the claim of Stansfield and Kenyon(1992) that the tape-recorded speaking test, in which there is no interlocutor , is 'fairer' than face-to-face speaking tests. Is the computer assisted test fairer than the face-to-face one? Savignon suggests communicative competence 'depends on the co-operation of all the participants involved' (1983:9). And part of the communicative competence is in knowing how to keep the conversation going, which includes knowing when to feign understanding and when to change the subject (Gunn, 2003). With no interlocutor involved in the computer assisted test, the issue of fairness and the capacity of items to test aspects of communicative competence will be will be important targets for data collection and analysis in this research.

The tape-based testing only covers some aspects of interactive speaking and the construct is more clearly connected with spoken production (Luoma, 2004). As with the tape-recorded test, the computer assisted test assesses only the spoken production of the testee rather than the interaction between testee and interlocutor found in interviews, role plays and other tests of speaking involving multiple speakers. The advantage of a computer assisted test is that the aural and visual stimuli remain precisely the same for all testees and, given the impersonality of the test procedure, differences due to inter-personal factors will be minimized. A question may occur in the test as to whether the response to such inauthentic stimuli can be regarded as authentic speech. Some believe that a face-to-face interview is most authentic because it is interactive. Underhill (1987) thought the voice-recorded test was not very authentic because the assessor of a recorded test can hear everything a live assessor can, but she cannot see the test, she therefore misses all the visual aspects of communication such as gesture and facial expression. A crucial question defining authenticity is 'authentic to what?' (Messick, 1994:18). Authenticity is not an objective quality as such; it is subjective and dependent on who is judging the authenticity (Gulikers, et al, 2006). There is almost no literature about authenticity of computer assisted speaking test; therefore in the research this issue will be investigated in detail.

Being afraid of poor performance in front of other people, students tend to be silent in class. This is particularly noticeable in Asian English as Second Language learning classes. In the 1990s related studies indicated that students who used to be shy in face-to-face discussion and who were considered low achievers in language learning became more active participants in computer-assisted classroom discussion (Beauvois, 1992, 1995; Kelm, 1992). Without seeing each other in the test, with a less

threatening means to communicate, students may find it easier to speak. In the computer-assisted tests, will the testees find it easier to speak, in the absence of visible testers? Will their language output be changed?

The essential challenge from advocates of face-to-face testing is that computer assisted assessment is not an authentic simulation of 'real life' language use. We would counter that face-to-face exchanges actually only represent on spoken language context. For students who may go on to an academic career, for example, we suggest that the ability to make a presentation – or give a lecture, if you will – to a large audience on a familiar topic may be a skill that will be required in the future. Furthermore, in this era of electronic communications, speaking over the telephone, or via an internet link, such as Skype, with or without visual contact, is something that will be a common part of these students' lives – perhaps more common than face-to-face encounters for many. But the argument is not so much one of which of these assessment contexts is 'more authentic' but rather that we should ask what forms of spoken language use any assessment best approximates to and may therefore for which it may claim some level of validity.

### **Human-computer interaction: who are you talking to?**

With rapid developments in computer-based technologies in recent years, the use of computers to administer tests is becoming increasingly common in education (Bonham et al, 2000; Mason et al, 2001; Olson, 2002). It is predicted that the use of computer-assisted tests for language assessment and other assessment purposes will become increasingly predominant in the immediate future (Bennett, 1999). However some researchers argue that these and other computer-linked factors may change the nature of a task so dramatically that one cannot say the computer-assisted and conventional version of a test are measuring the same thing (McKee & Levinson, 1990).

Changing the administration of the test may affect the reliability of a test. Computer-based test provides testees with an equal opportunity by allowing every testee to have the very same testing experience. Introducing a new method of assessment however may cause students anxiety. For many people, the test situation itself creates considerable anxiety which can badly affect their performance (Underhill, 1987). However, Foot (1999) highlights that students may not necessarily perform better if they are more relaxed. In general, higher-attaining students will adapt most quickly to any new assessment approach (Watson, 2001; Noyes, 2004) and will quickly develop test-taking strategies that benefit them in the new approach. Computer anxiety is another potential disadvantage that may affect test performance (Henning, 1991). Also differences in the degree to which students are familiar with using computers may lead to differences in their performances on computer-assisted or computer-adaptive tests (Hicks, 1989; Henning, 1991). Clark (1988) and Stansfield et al (1990) found that examinees sometimes felt nervous taking a computer assisted test, because of a feeling of lack of control. Some examinees reported that they felt this nervousness prevented them from doing their best.



Research into computer-supported learning suggests that women suffer from lower levels of computer literacy and lower confidence levels in its use (Yates, 2001). Men and women were also observed to behave differently in on-line group discussion (Barrett & Lally, 1999). In particular, it was observed that men's talk was, typically, more numerous and longer than that of women, and tended to include greater levels of social exchange. Women, however, appeared, typically, to be more interactive than men. Some studies claim that the internet increases engagement, confidence, and responsibility with a less threatening means to communicate (Chun, 1994; Beauvois, 1995; Skinner & Austin, 1999), while McGrath (1997-98) found that those students who do well in a face-to-face environment may be suppressed in a web-based environment and vice versa. There is a large body of research in the field of gender, familiarity and anxiety on human & computer interaction, while almost no research has been done on comparing differences in behaviour and speech when human beings are speaking to a computer rather than to other human beings.

### **Towards a Research Agenda**

It is clear from the discussion above that a full evaluation of CEOTS, even just in relation to its possible use in the CET system, demands the investigation of many factors. We further recognise, however, that given the very recent appearance of computer assisted assessment of spoken language and the apparent shortage of research into speech based interactions with computers this is an opportunity to carry out more fundamental research that goes beyond an evaluation of a particular system.

As suggested above, we believe that the starting point for our evaluative research is to ask what interpretations can be validly made of performance in any given form of spoken language assessment, rather than to start with a notion of what is 'authentic' or 'non-authentic'. This suggests to us that one of our chief research tasks is to analyse the spoken language generated under various testing circumstances and by different tasks set within those circumstances. We are fortunate in having a huge volume of test results – including the actual voice recordings – available to us through our collaboration with USTC. We also have the interest and co-operation of the CET administration in this project and through them will have access to video recordings of a large number of their face-to-face tests. These clearly present opportunities for detailed linguistic analysis of the responses generated by different item formats and individual items in the test, for some of which we shall use linguistic analysis computer software. The precise nature of the aspects of language we shall be looking for remain to be firmly established but we hope that we shall be able to produce a 'profile' of language responses to testing modes and item types.

Despite the existence of this considerable database, however, we feel there is a need to collect data under more controlled conditions. We are in particular interested in investigating testee responses beyond the linguistic and plan to video individuals taking the computer assisted tests to allow us to analyse

face and body activity. We would also like to examine their subjective perceptions of the two forms of testing, which we shall do through both quantitative survey techniques and using individual and group interviews to obtain data for qualitative analysis.

Interviews will also be held with English teachers at USTC, particularly those who have experience of teaching before and after the introduction of CEOTS, to obtain –admittedly somewhat subjective – data on the impact that the testing has had on their language classes.

Data on aspects of the reliability of the speaking test results will be collected in a variety of ways. Test-retest reliability data will be generated for selected groups of students. The grading process will be subject to scrutiny by observation of the process, through interviews with markers and by comparing marks from different markers for the same recordings. Some similar data will be collected for face-to-face tests, and it is hoped that the co-operation that the national CET committee have promised us will give us access to their own data on the reliability of the CET-SET.

Finally, issues of manageability of the system, particularly with respect to a potential huge increase in scale, will be examined through discussions with USTC staff involved in the system management and development.

## **Conclusions**

Computer assisted speaking testing is a relatively new field and there are, as yet, few large and detailed studies in this field. Using computers can potentially allow simultaneous performance of the speech production part of testing by a large number of students, although the grading of their performances remains labour-intensive. Whether a system such as CEOTS can overcome the inefficiency of traditional face-to-face testing and make oral testing on a large scale possible, without major detrimental impact on the validity and other aspects of the assessment, is at the heart of this research. We recognise the complexity of the project that we are taking on and anticipate that we shall be continually reviewing both our methodological and theoretical approaches. We remain convinced, however, that alternatives to face-to-face testing must be found so that the annual ten million plus English language testing candidates in Chinese universities can be offered a test of their speaking competence. If this is not done, then the backwash effect of the absence of such an examination for the vast majority will continue to distort the teaching and learning of English among those students and to undermine the government's attempts to improve the language proficiency of the country's university graduates. The research in which we are currently engaging promises therefore to be of considerable practical and theoretical significance.

## References

American Psychological Association (APA), American Educational Research Association (AERA) and National Council on Measurement in Education (NCME) (1966) *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

Argyle, M. (1983) *The Psychology of Interpersonal Behavior* (4th ed). Harmondsworth: Penguin.

Bachman, L. and Palmer, A. (1996) *Language Testing in Practice*. Oxford: Oxford University Press.

Barrett, E. and Lally, V. (1999) Gender differences in an on-line learning environment. In *Journal of Computer Assisted Learning*, Vol. 15, pp48-60.

Beauvois, M. H. (1992) Computer-assisted classroom discussion in the foreign language classroom: Conversation in slow motion. In *Foreign Language Annals*, Vol. 25(5), pp 455-464.

Beauvois, M. H. (1995) E-talk attitudes and motivation in computer-assisted classroom discussion. In *Computer and the Humanities*, Vol. 28, pp177-190.

Bennett, R. E. (1999) Using new technology to improve assessment, In *Educational measurement: Issues and practice*, Vol. 18(3), pp5-12.

Black, P. (1998) Assessment and classroom learning. In *Assessment in Education*, Vol. 5, pp7-74.

Bonham, S. W., Beichner, R. J., Titus, A., and Martin, L. (2000) Education research using Web-based assessment systems. In *Journal of Research on Computer in Education*, Vol. 33, pp28-45.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. In *Applied Linguistics*, Vol. 1(1), pp8-24.

Chun, D. (1994) Using computer networking to facilitate the acquisition of interactive competence. In *System*, Vol. 22, pp17-31.

Clark, J. L. D. (1988). Validation of a tape-assisted ACTFL/ILR-scale based test of Chinese speaking proficiency. In *Language Testing*, Vol. 5(2), pp197-205.

Foot, M. C. (1999) Relaxing in Pairs. In *ELT Journal*, Vol. 53(1), pp36–41.

Gipps, V. C. (1994) *Beyond testing-towards a theory of educational assessment*. Washington, D.C.: The Falmer Press.

Goodwin, C. (1981) *Conversational organization: interaction between speakers and hearers*. New York: Academic Press.

Gulikers, J., Bastiaens, T. and Kirschner, P. (2006) Authentic assessment, student and teacher perceptions: the practical value of the five-dimensional framework. In *Journal of Vocational Education and Training*, Vol. 58(3), pp337-357.

Gunn, L. C. (2003) Exploring second language communicative competence. In *Language Teaching Research*, Vol. 7(2), pp240-258.

Heaton, J. (1988) *Writing English Language Tests*. London: Longman

Henning, G. (1991) Validating an item bank in a computer-assisted or computer-adaptive test. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (New York: Newbury House), pp209-222.

Hicks, M. (1989) The TOEFL computerized placement test: Adaptive conventional measurement. *TOEFL Research Report No. 31*. Princeton, NJ: Educational Testing Service.

Hughes, A. (1989) *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Jin, Y. and Yang, H. Z. (2006) The English proficiency of college and university students in China: as reflected in the CET. In *Language, Culture and Curriculum*, Vol. 19(1), pp21-36.

Kelm, O. R. (1992) The use of synchronous networks in second language instruction: a preliminary report. In *Foreign Language Annals*, Vol. 25, pp441-545.

Kraut, R.E., Fish, R.S., Root, R.W. and Chalfonte, B.L. (1990) Informal Communication in Organizations: Form, Functions and Technology. In: Oskamp, S. and Spacapan, S. (Eds.) *Human Reactions to Technology. The Claremont Symposium on Applied Social Psychology*. Beverly Hills, CA: Sage Publication.

Kyriakides, L. (2004) Investigating validity from teachers' perspectives through their engagement in large-scale assessment: the emergent literacy baseline assessment project. In *Assessment in Education*, Vol. 11(2).

Luoma, S. (2004) *Assessing speaking*. Cambridge: Cambridge University Press

Madaus, G. and Pullin, D. (1991) To audit and validate 'high stakes' testing programs, in: R. G.. O'Sullivan (Ed.) *Advances in program evaluation*: Vol. 1A,

Effects of mandated assessment on. teaching (Greenwich, CT, JAI Press), pp139-158.

Mason, B. J., Patry, M., and Bernstein, D.J. (2001) An examination of the equivalence between non-adaptive computer-based test and traditional testing. In *Journal of Educational Computing Research*, Vol. 24, pp29-39.

McGrath, C. (1997-98) A new voice on interchange: is it talking or writing? Implications for the teaching of literature. In *Journal of Educational Technology systems*, Vol. 26(4), pp291-297.

McKee, L. M. and Levinson, E. M. (1990) A review of the computerized version of the self-directed search. In *Career Development Quarterly*, Vol. 38(4), pp325-333.

Messick, S. (1989) Validity. in R. L. Linn (Ed.), *Educational measurement (3rd ed.)*. New York: American Council on Education & Macmillan.

Messick, S. (1994) The interplay of evidence and consequences in the validation of performance assessment, in *Educational Research*, Vol. 23(2), pp13-23.

Noyes, J., Garland, K. and Robbins L. (2004) Paper-based versus computer-based assessment: is workload another test-mode effect? In *British Journal of Educational Technology*, Vol.35 (1), pp111-113.

O' Loughlin, K. (2002) The impact of gender in oral proficiency testing. In *Language Testing*, Vol. 19(2), pp169-192.

Olson, A. (2002) Technology solution for testing. In *School Administration*, Vol. 59, pp20-23.

Read, J. and Chapelle, C. A. (2001) A framework for second language vocabulary assessment. In *Language Testing*, Vol. 18(1), pp1-32.

Roever, C. and McNamara, T. (2006) Language testing: the social dimension. In *International Journal of Applied Linguistics*, Vol. 16(2).

Savignon, S. (1983) *Communicative competence: Theory and classroom practice*. Reading, MA: Addison-Wesley

Seddon, G. M. and Pedrosa, M. A. (1990) Non-Verbal Effects in Oral Testing. In *British Educational Research Journal*, Vol. 16(3), pp305-310.

Skinner, B. and Austin, R. (1999) Computer conferencing: Does it motivate EFL students? In *ELT Journal*, Vol. 52(1), pp38-42.

Stansfield, C. W., Kenyon, D. M., Paiva, R., Doyle, y F., Ulsh, I., and Cowles, M. A. (1990). The development and validation of the Portuguese Speaking Test. In *Hispania*, Vol. 72, pp641-651.

Stansfield, C.W., & Kenyon, D.M. (1992) The development and validation of a simulated oral proficiency interview. In *Modern Language Journal*, Vol. 76, pp 129-141.

Underhill, N. (1987) *Testing spoken language*. Cambridge: Cambridge University Press.

Waston, B. (2001) Key factors affecting conceptual gains from CAL. In *British Journal of Educational Technology*, Vol. 32(5), pp587-593.

Wolf, R. M. (1998) Validity issues in international assessments. In *International Journal of Educational Research*, Vol. 29(5), pp491-501.

Worthen, B. R., Borg, W. R. and White, K. R. (1993) *Measurement & evaluation in the schools*. London: Longman.

Yates, S. J. (2001) Gender, Language and CMC for education. In *learning and Instruction*, Vol. 11, pp21-34.