INNOVATIVE ASSESSMENT AND PERSONALISED FEEDBACK IN HIGHER EDUCATION

Jill Barber and Steven Ellis
University of Manchester
United Kingdom

Abstract

Since adopting summative on-line examinations in 2005, we have increased the range of question types to include short essays and questions incorporating chemical structures and now achieve time savings of up to 90% in the marking process. Online assessments allow two novel forms of feedback: (a) an anonymised spreadsheet containing all the marked exam scripts is made available to all students and (b) *Smallvoic*e, a novel app, provides confidential personalised feedback. Feedback statements, though written by the instructor, are selected by a computer in response to various aspects of a student's performance. There is evidence of improved student satisfaction and improved learning.

Introduction

Decades after the introduction of online assessment, it is still practised by a small minority of higher education institutions (Bull & McKenna, 2004). Nobody doubts that online assessment is possible –we all take part in surveys powered by such apps as Survey Monkey - but there is widespread doubt that it is secure, flexible or reliable.

An online assessment must of necessity be mounted on a server ahead of the examination period; this means that it can, in principle, be hacked. There is a body of literature devoted to the security of online assessments (Apampa, Wills, & Argles, 2009). Of course, it is, in principle possible for students to break into a university safe containing paper-based examinations, but this is a familiar risk, one that we are content to live with. Even though university online security systems are generally of a very high standard (universities maintain personal and often medical data), it is much easier for academic staff to imagine their students as computer hackers than as safe breakers.

It is still widely believed that a computer based assessment is limited to multiple choice and related question types. Of course, these types of question are very valuable and can be automatically marked, but even the most rudimentary assessment software allows for free text entry, permitting short answers and even essays to be submitted and marked online. There are even interfaces that permit submission of drawings.

Reliability is, however, most academics' major concern around e-assessment. Computers are inherently unreliable. A typical 50-seat computer cluster in a university might be expected to have two or three computers out of commission for various reasons at any one time. This is a level of unreliability we would find unacceptable in our cars or washing machines. The fear is always of a computer failure mid-assessment, so that student work is lost. Some of the published literature incorporates elaborate schemes for backing up student work on paper (Aojula, Barber, Cullen, & Andrews, 2006).

Further, an online assessment produces dependence. An academic conducting a paper-based assessment maintains the impression or illusion of control. The photocopied examination papers are retrieved from a safe shortly before the examination, and physically handed out to students. Invigilation requires no very special expertise; we understand what is going on. Academics conducting online examinations become dependent upon IT staff, whose expertise is usually quite alien. Attempts to guide academics through the process serve to reinforce dependence (Willis et al., 2009).

All these impressions of online assessment are uncomfortable, and it is all too easy to defer the transition to online assessment.

Security, Flexibility and Reliability: The Manchester Experience 1998 – 2004

In 1997 we began a project entitled "What makes a student succeed?" (Sharif Gifford, Morris, & Barber, 2003, 2007a, 2007b), requiring the use of several assessments in the first week of the MPharm course. The results were used to assign students to foundation groups, and the assessments were therefore high stakes, though not summative. The start of this project coincided with the university's introduction of so-called CBA (computer-based assessment) software, which was a modification of the commercially available Questionmark software.

Security 1998-2004

Students sat the password-protected tests in university computer clusters. Passwords were issued immediately prior to the tests, which were invigilated and conducted under standard examination conditions (no books, paper, coats, etc., permitted). There is no evidence of any security breach at any point.

Flexibility 1998-2004

Question types were limited to automatically-marked multiple choice, text match and numerical questions. While this was adequate for the purpose of assigning students to foundation groups, it would not permit a full range of assessments in a Pharmacy programme to be conducted online.

Reliability 1998-2004

During this period, all our worst fears about computer reliability were realised. The testing proceeded smoothly only in 2002. In every other year there was a failure of some sort.

- **1998** Many students were unable to register on the university IT system, were unable to access the online tests and sat paper versions.
- **1999** Guest logins were developed to solve the previous year's problem, but the traffic on the university network was so great at the start of the semester that login was unacceptably slow.

- **2000** The CBA servers were replaced and logins were staggered; unfortunately a scheduled change in the university computer image resulted in a widespread crash, affecting half the computer cluster.
- **2001** The university suffered a major virus attack, and the tests were conducted on paper.
- **2003** While everything went well at the point of testing, there was an error in the recording of results by the software, and it proved impossible to assign the marked tests to the correct students.
- **2004** The computer cluster was housed in a 19th century building in a city known for its rainfall – a flood in the cluster resulted in the tests again being carried out on paper.

In 2004 it was clear that the support for CBA software was inadequate for summative assessment.

Security, Flexibility and Reliability: The Manchester Experience 2005 – 2012

In 2005, we began to explore the use of WebCT (later to be superseded by Blackboard 8 and Blackboard 9) for summative assessment. This was prompted by an increase in student numbers in a first year Cell Biology and Biochemistry class to over 200. The teaching time on the unit was 48 hours, but the paper-based assessment was taking 60 hours to mark. The paper-based assessment was replaced by an online examination in which the questions were all automatically marked, a mixture of multiple choice text match questions. Human intervention in the marking process was now minimal, with less than one hour required.

Security 2006 – 2012

We were satisfied that students would not be able to access the assessment prior to the examination, but at the beginning of this period we were concerned that students might be able to access unauthorised websites during the assessment. Invigilators were trained to check the computer taskbars for minimised icons and to investigate any that looked suspicious. The architecture of the clusters also gave cause for concern – these had been designed for teaching and private study, with collaboration encouraged. Invigilators also had to be wary of students looking at one another's screens.

Flexibility 2006 – 2012

During this period, we gradually increased the range of question types available. Short answers and even short essays are supported by the Blackboard 9 assessment software. These can be marked either online within Blackboard or by downloading a csv file and marking in an Excel spreadsheet. Marking in a spreadsheet is undoubtedly quicker than marking online because students' answers to a particular question are arranged in a column; scrolling from one answer to the next is much quicker than closing one file and opening another.

We developed a bank of chemical structures such that the answer to a question could be a chemical structure. Then 153 structures were classified according to the heteroatoms (atoms other than C, H, N, O) they contain, the number and size of rings and functional groups. Filters allowed students to select appropriate structures quickly (see Figure 1).

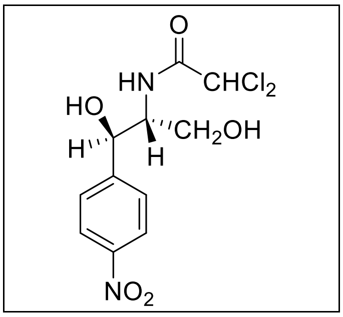


Figure 1. Chloramphenicol: the structure contains two chlorine atoms, one six-membered ring and an amide function.

Reliability 2005 – 2012

Both the university IT systems and Blackboard proved very reliable during the period up to 2012. We developed various pieces of bespoke software that allowed us, for example, to upload examination papers directly from Word, incorporating diagrams and chemical structures.

During the January 2013 examination period, several examinations were hit by a failure to save answers, affecting about 17% of students. This was tracked to a failure in the six Blackboard servers to transfer load effectively when traffic was high. The problem took several weeks to identify, at which point it was quickly corrected.

Discussion

Security of online examinations remains a concern. Nevertheless, this is a concern that universities have so far proved able to meet. As of 2013, online examinations did not provide the flexibility of question type that we ultimately require, but neither were they restricted to multiple choice questions. Reliability remains the key issue that prevents many academics embracing online assessment (Warburton, 2009). Online examinations present a new range of challenges and when something goes wrong, it is seldom the same as what went wrong in the previous year. As a consequence, academic staff lose control of the examination, and this is unnerving.

The Advantages of Online Assessment: Accuracy, Speed and Feedback

The 2013 failure resulted in a significant loss in confidence in online assessment, and a reduction in the number of examinations conducted online. It prompted a reconsideration of the advantages of online assessment, as well as the development of additional security features.

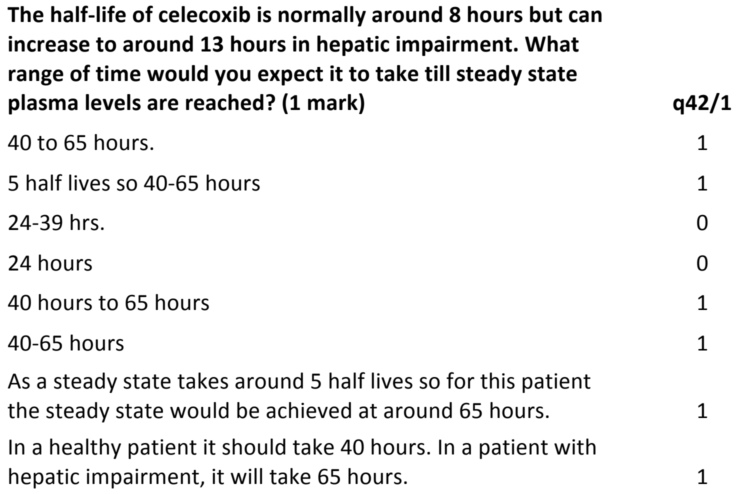
Accuracy

The advantages of online assessment have been clearly outlined by McGee Wood, Sargeant, & Jones, (2005) and latterly Briggs (2015). The first and indisputable advantage is that computers are good at adding up – they total marks awarded quickly and accurately. Although academics like to think they are good at adding up, the evidence (Aojula et al., 2006) is that they are not. Andrews (2005, personal communication) demonstrated that the error rate is typically 5% in a moderately complex assessment. It is therefore imperative that an assessment that is marked manually be totalled at least twice.

Speed (and accuracy)

Nobody disputes that computers write legibly. A second advantage of online text-based examinations is that students' handwriting is often difficult to decipher. It is much faster marking typescript, which is inherently legible. McCann (2010) has demonstrated that these obvious advantages may not be sufficient to persuade academics to invest the initial effort in mastering the logistics of e-assessment.

The real insight made by McGee Wood et al. (2005) is that computers are good at finding script. Figure 2 shows a question, answered by many students, with the answers arranged one above the other in a spreadsheet.



The half-life of celecoxib is normally around 8 hours but can increase to around 13 hours in hepatic impairment. What range of time would you expect it to take till steady state plasma levels are reached? (1 mark)	**q42/1**
40 to 65 hours.	1
5 half lives so 40-65 hours	1
24-39 hrs.	0
24 hours	0
40 hours to 65 hours	1
40-65 hours	1
As a steady state takes around 5 half lives so for this patient the steady state would be achieved at around 65 hours.	1
In a healthy patient it should take 40 hours. In a patient with hepatic impairment, it will take 65 hours.	1

Figure 2. Example of an online examination question and its marking.

When answers are arranged in this way, there is no time spent rifling through answer books trying to find the answer to a particular question. Academics who mark on a spreadsheet normally estimate time savings of a factor of between 2 and 10, depending upon the length of the typical answer (the time savings are less with longer answers where more time is spent marking and less is spent finding the answer). That most of the time saving arises from finding the script quickly is evidenced by a direct comparison. In Blackboard, it is possible to download the students' answers as a csv file, but it is also possible to mark directly onto an online document resembling an examination script. The former is much faster, perhaps by a factor of 2.

Less immediately obvious is the fact that the same spreadsheet format permits improvements in accuracy of marking. After marking a question, it is easy to sort the spreadsheet by mark and to confirm that answers achieving the same mark are comparable. This can almost never be achieved in a paper-based examination.

Feedback

Marking on a spreadsheet leads to such confidence in consistency of marking that complete transparency is possible. Beginning in 2008, after securing the permission of the students, we stripped all identifiers from the marked examination spreadsheet and made it available to all the students who had sat the examination; essentially students saw Figure 2, but extended for all questions. Thus students could not only see where they gained or lost marks, but could see a range of very good answers for each question.

Students are often unconvinced that time savings for staff (even the massive savings afforded by online assessment) are of any benefit to them. They do not accept that time saved marking will be fed into teaching. They do, however, appreciate the feedback, and focus groups return to this point frequently.

More Feedback – The Smallvoice App

It is quite common for a student to approach an academic following an assessment and to say, “Where did I go wrong?” Armed with an online assessment, the academic may scan the row corresponding to the particular student’s assessment and offer some analysis.

Typical comments include:

- You haven’t answered all the questions.
- You haven’t revised a particular topic
- Your English is poor

Very often, even usually, such analysis could just as well be given by a well-programmed computer, as in the examples above. We therefore developed Smallvoice.

Smallvoice provides rapid, automated, completely personal feedback on performance to students in large classes. It analyses many different aspects of a student’s performance and synthesises accurate, confidential advice. Smallvoice is a freestanding tool, able to integrate with commonly-used data systems around the world.

Smallvoice analyses an examination paper (either a computer-based examination paper or a transcript of marks from a paper-based examination) in the same way as an instructor might analyse a paper following an examination. It reports on a student’s performance in different topics (for example different diseases), different question types (e.g., factual recall, multiple choice, critical argument). In addition it analyses performance in ways that are much easier for a computer program than for an instructor. It incorporates a powerful algorithm for discrimination values, so is able to comment on whether a student fared batter in the easier or more difficult questions relative to the rest of the class. It correlates performance in summative assessment with attendance and with performance in past formative and summative assessments. Students receive a detailed email showing where they are in the class, trends in their performance, and incorporating links to sophisticated statistics about individual questions. The feedback is made up of text inputted by the instructors and is therefore personal in tone; it is at its most powerful when used to congratulate good students, to encourage average and weaker students and to give advice about preparation for future learning.

Smallvoice hugely increases student satisfaction. We have received numerous emails of appreciation from students, and a Feedback score of 4.69 /5 in the university's course evaluation questionnaire in the pilot course unit. We have also seen average marks rise 10-20% between successive examinations following feedback. Smallvoice lends itself to feedback that advises students about improving performance, which (like the personal tone) is a hallmark of current perceptions of good feedback (Price, Handley, Millar, & O'Donovan, 2010; Boud & Malloy 2013).

Sample Output

This is an example of part of the feedback used to support the end of semester one examinations for fourth year students. Smallvoice can also be used to give very fine-grained feedback (for example a discussion of individual questions in a single course assessment.

Dear [Forename],
Here is some feedback following your semester one exams. Your weighted mean for semester 1 was 65.4% and the mean for the cohort was 67.2%. Your position in the group was therefore 98=. This was a good solid 2.1 performance in semester one. Well done! Your semester 1 mark is significantly higher than your year 3 mark so very well done!

The second year contributes 10% to your final degree classification and the third year contributes 20%. In the fourth year so far you have completed 50 credits out of 120, that's another 29.2% of your degree. 40.8% remains.

Table 1 shows you the average mark you need in semester 2 to get each class of degree.

Table 1

Averages required for the rest of your degree

to get a first you need	to get a 2.1 you need	to get a 2.2 you need	to get a third you need
80.1%	55.7%	31.2%	6.7%

Do remember though, that the average is not quite the whole story. You have to pass all the modules!

Table 2 shows a summary of your module marks compared with the class averages. Your mark in Law was especially commendable.

Table 2:
Summary of your module marks

Module	Law	Dispensing	Social Pharmacy	Microbiology	Neuropharmacol
Your Mark	80.9	79.2	63	68	58
Your Position	47=	112=	128=	64=	20
Number in class	170	170	170	152	29
Class mean mark	75.2	81.5	69.9	64.8	61.0

You're progressing very well. Good luck with the rest of the semester.
Best wishes
Jill and Steve

The Future of Online Assessment

Given the advantages of online assessment to both academic staff and students, progress in delivering secure, flexible, easily-managed and (above all) reliable assessment has been disappointing. The delivery of online assessment requires an enormous amount of care, and the support of local in-house IT experts.

Examination infrastructure
To ensure security during examinations, the University of Manchester has developed computer clusters specifically for examinations. Computers are widely spaced and screens cannot easily be seen by a student's neighbours. A specific *examination desktop* is loaded onto the cluster machines prior to the examination period and websites outside Blackboard cannot be viewed.

This feature has led to the development of a novel online open-book examination format, in which students are able to access specific materials contained within the same Blackboard folder as the examination.

The conduct of online examinations is now coordinated by a specific member of the Examinations Office. Protocols for paper-based examinations have evolved over many decades to accommodate several examinations taking place in the same very large room. Online examinations present a new paradigm. A single examination may be housed in several different remote rooms. Ensuring consistency between rooms is a significant challenge, requiring efficient communication between several rooms (carried out via online messaging).

Load testing
The 2013 failure prompted us to develop load testing protocols to be carried out ahead of every examination period. The intention during load testing was to provide evidence that the current deployment of our virtual learning

environment was fit for purpose and that there was a relatively low risk of encountering any load related issues during the setup or running of our online exams. Several clusters of desktop machines were used in the testing with a combined provision of approximately 400 machines. A version of the Mozilla Firefox browser was modified so that it could simulate individual student activity during setup and running of an online examination. This browser was started on each machine so that the behaviour of 400 virtual students could be arranged and synchronised during the period allocated for testing.

Two tests are conducted on the Blackboard infrastructure. The first introduces gradual load (achieved by conducting a real exam on each PC) onto Blackboard up to the maximum PCs available across all the clusters used. When this capacity is reached, the exam is allowed to continue for approximately 15 minutes to test for sustained load on Blackboard. The second test starts all the exams together to simulate peak load of the system.

The virtual learning environment configuration at the University of Manchester is currently composed of 10 application servers. The advice from our hosting partners has been that this configuration is over specified for our actual use. The intention of load testing was to prove that the servers would cope without failure with the load being generated during examinations and equally important that they would comfortably do so. Whilst it is often difficult to correlate load and system utilisation, one measure that can be used is the number of queued processes within the processor of an application server. The larger the number the more likely there is to be service degradation, or service loss (either partial or full). If a sustained load of "20" was observed, an investigation was triggered. If an application server reached a value of "50," an automated procedure would take it out of the processing pool so that no additional load would be transferred to it. During both tests undertaken during our recent load testing, the maximum number of queued processes observed was "6," with a typical value being between "1 and 3."

Load testing is, we believe a necessary prelude to online examinations.

Downloads

Downloading examinations in csv format also requires specialist tools. Blackboard, for example, enables html, which is not rendered directly. In general, this is removed manually.

More inconvenient still is that students occasionally use a character, such as a hyphen, as a bullet point. Excel recognises this as a delimiter, and a student's answer may be truncated as a result of its use. The solution is to brace each answer inside | characters, which can be achieved in a number of ways, by opening the csv file initially in a program other than Excel.

Drawing tools

In Pharmacy and related subjects, online assessment will only come of age when drawing diagrams and chemical structures within the assessment are enabled.

Discussion

Online assessment has the potential to be enormously powerful, saving time, giving improved accuracy and transparency and greatly facilitating feedback. Holmes (2015) and others have also pointed to the frequent use of simpler e-assessments as a means of improving student engagement.

Despite nearly two decades of experimentation, however, it is still not for the faint-hearted. Commercial packages require sophisticated in-house support and supplementation if they are to be used reliably at high volume. Accuracy, speed and feedback are still achieved at the expense of security, flexibility and reliability.

References

Aojula, H., Barber, J., Cullen, R., & Andrews, J. (2006). Computer-based, online summative assessment in undergraduate pharmacy teaching: The Manchester experience. *Pharmacy Education, 6*, 229–236.

Apampa, K.M., Wills, G., & Argles, D. (2009) Towards security goals in summative e-assessment security. Internet technology and secured transactions. Retrieved from ieeexplre.ieee.org.

Briggs, N. (2015). *Summative online assessments outside the examination period* (Internal working paper). Manchester, UK: University of Manchester.

Boud, J., & Molloy, E. (2013). Rethinking models of feedback for learning: the challenge of design. *Assessment & Evaluation in Higher Education 38*, 698–712.

Bull, J., & McKenna, C. (2004). *Blueprint for computer-assisted assessment*. London: Routledge Farmer.

Holmes, N. (2015). Student perceptions of their learning and engagement in response to the use of a continuous e-assessment in an undergraduate module. *Assessment & Evaluation in Higher Education, 40*, 1–14.

McCann, A. (2010). Factors affecting the adoption of an e-assessment system. *Assessment & Evaluation in Higher Education 35*, 799–818.

McGee Wood, M., Sargeant, J., & Jones, C. (2005). What students really say. *Proceedings of the 9th CAA Conference.* Loughborough, UK: Loughborough University. Retrieved from http://caaconference.co.uk/pastConferences/2005.

Price, M., Handley, K., Millar, J., & O'Donovan, B. (2010). Feedback: All that effort, but what is the effect? *Assessment & Evaluation in Higher Education, 35*, 277–289.

Sharif, S., Gifford, L., Morris, G. A., & Barber, J. (2003). Can we predict student success (and reduce student failure)? *Pharmacy Education, 3*, 77-86.

Sharif, S., Gifford, L., Morris, G.A., & Barber, J. (2007a). Diagnostic testing of first year pharmacy students: A tool for targeted student support. *Pharmacy Education, 7*, 215-221.

Sharif, S., Gifford, L., Morris, G.A., & Barber, J. (2007b). An investigation of the self-evaluation skills of first year pharmacy students. *Pharmacy Education, 7*, 295-302.

Warburton, B. (2009) Quick win or slow burn: Modelling UK HE CAA uptake. *Assessment & Evaluation in Higher Education 34*, 257–272.

Willis, G. B., Bailey, C P., Davis, H. C., Gilbert, L., Howard, Y., Jeyes, S., Millard, D. E., Price, J., Sclater, N., Sherratt, R., Tulloch, I. , & Young, R. (2009). An e-learning framework for assessment (FREMA). *Assessment & Evaluation in Higher Education, 34*, 273–292.

Author Details

Jill Barber
Jill.barber@manchester.ac.uk

Steven Ellis
Steve.ellis@manchester.ac.uk