

E-assessment: Past, present and future

Sally Jordan

Department of Physical Sciences, The Open University

Abstract

This review of e-assessment takes a broad definition, including any use of a computer in assessment, whilst focusing on computer-marked assessment. Drivers include increased variety of assessed tasks and the provision of instantaneous feedback, as well as increased objectivity and resource saving. From the early use of multiple-choice questions and machine-readable forms, computer-marked assessment has developed to encompass sophisticated online systems, which may incorporate interoperability and be used in students' own homes. Systems have been developed by universities, companies and as part of virtual learning environments.

Some of the disadvantages of selected-response question types can be alleviated by techniques such as confidence-based marking. The use of electronic response systems ('clickers') in classrooms can be effective, especially when coupled with peer discussion. Student authoring of questions can also encourage dialogue around learning.

More sophisticated computer-marked assessment systems have enabled mathematical questions to be broken down into steps and have provided targeted and increasing feedback. Systems that use computer algebra and provide answer matching for short-answer questions are discussed.

Computer-adaptive tests use a student's response to previous questions to alter the subsequent form of the test. More generally, e-assessment includes the use of peer-assessment and assessed e-portfolios, blogs, wikis and forums.

Predictions for the future include the use of e-assessment in MOOCs (massive open online courses); the use of learning analytics; a blurring of the boundaries between teaching, assessment and learning; and the use of e-assessment to free

Corresponding author:

Sally Jordan, Department of Physical Sciences, The Open University, UK
Email: sally.jordan@open.ac.uk

human markers to assess what they can assess more authentically.

Keywords: E-assessment, computer-marked assessment: review

Introduction

E-assessment, according to its widest definition (JISC 2006), includes any use of a computer as part of any assessment-related activity, be that summative, formative or diagnostic. So its scope includes the online submission of an assignment for marking by a human, the assessment of an e-portfolio or reflective blog, feedback delivered by audio files recorded on a computer and, most commonly, online computer-marked quizzes. Other terms with similar meaning include technology-enhanced or technology-enabled assessment and computer-assisted or computer-aided assessment.

Some of the recent reviews of e-assessment literature (e.g. Conole & Warburton 2005, Dikli 2006, Hepplestone *et al.* 2011, JISC 2009, Kay & LeSage 2009, Nicol 2008, Ridgway *et al.* 2004, Ripley 2007, Stödborg 2012) have focused on a subset of technology-enhanced assessment or feedback. This review is deliberately broad, although its main focus is on computer-marked assessment. Even this is a huge field, so the review is of necessity selective, mentioning only a few of the hundreds of papers referring to particular technologies (e.g. electronic voting systems or 'clickers'). It has not been possible to mention all e-assessment systems (for a more comprehensive list, see Crisp (2007) pp69–74) are given as exemplars. Where choices have been made as to which systems and papers to include, those that have been developed, used and/or evaluated in the context of physical science or related disciplines have been favoured.

Drivers and development

"Effective assessment and feedback can be defined as practice that equips learners to study and perform to their best advantage in the complex disciplinary fields of their choice, and to progress with confidence and skill as lifelong learners, without adding to the assessment burden on academic staff. Technology . . . offers considerable potential for the achievement of these aims."

(JISC 2010, p8)

E-assessment is a natural partner to e-learning (Mackenzie 2003) offering alignment of teaching and assessment methods (Ashton & Thomas 2006,

Gipps 2005). It offers increased variety and authenticity in the design of assignments and, for example by means of e-portfolios, simulations and interactive games, it enables the assessment of skills that are not easily assessed by other means (JISC 2010).

Even the simplest of multiple-choice quizzes can enable students to check their understanding of a wide range of topics, whenever and wherever they choose to do so (Bull & McKenna 2004). Students thus have repeated opportunities for practice (Bull & Danson 2004), sometimes with different variants of the questions (Jordan 2011). Feedback can be provided instantaneously and can be tailored to particular misunderstandings, with reference to relevant module materials (Jordan & Butcher 2010). This provides, even for students who are studying at a distance, a virtual 'tutor at the student's elbow' (Ross *et al.* 2006). E-assessment allows students to make mistakes in private (Miller 2008) and the feedback is perceived to be impersonal (Earley 1988) and non-judgemental (Beevers *et al.* 2010).

Regular online tests have been shown to improve performance on an end of year examination (Angus & Watson 2009). Online assessment can engage and motivate students (Grebenik & Rust 2002, Jordan 2011) and help them to pace their study. Students can use the online assignments to check their understanding and so to target future study, but the mere act of taking tests has been shown to improve subsequent performance more than additional study of the material, even when tests are given without feedback. This is the so-called testing effect and research in this area is reviewed in Roediger & Karpicke (2006).

E-assessment thus has much to offer in terms of improving the student learning experience. However, it is interesting to note that the term 'objective questions', used to describe multiple-choice questions in particular, reflects the fact that the early use of multiple-choice came from a desire to make assessment more objective. The earliest multiple-choice tests were probably E.L. Thorndike's Alpha and Beta Tests, used to assess recruits for service in the US Army in the First World War (Mathews 2006). However, multiple-choice testing as an educational tool gained in popularity during the 20th Century as researchers became more aware of the limitations of essays (D.R. Bacon 2003). Ashburn (1938) noted a worrying variation in the grading of essays by different markers, an oft-repeated finding (e.g. Millar 2005). Human markers are inherently inconsistent and can also be influenced by their expectations of individual students (Orrell 2008). Multiple-choice questions bring objectivity whilst computer-marking brings a

consistency that can never be assured between markers or, over time, for the same human marker (Bull & McKenna 2004, Butcher & Jordan 2010).

Alongside increased reliability, computer-marked assessment can bring savings of time and resources (Dermo 2007), although writing high-quality questions should not be seen as a trivial task (Bull & McKenna 2000). Computer-marking is particularly useful for large class sizes (Whitelock & Brasher 2006) and it can add value and enable practitioners to make more productive use of their time (JISC 2010).

During the 20th century, large-scale multiple-choice tests were administered by means of machine-readable forms of the type shown in Figure 1, on which students indicated their selected answer to each question. These systems (which are still in use) enabled objectivity and resource saving, but the advantages of immediacy of feedback and student engagement were not yet present. Other question types existed, but at the time of Brown *et al.*'s (1999) review of practice in higher education, e-assessment was essentially synonymous with 'multiple-choice', with the authors concluding that assignments simply converted from paper to screen usually proved inadequate.

Since around 1980, there has been a rapid growth in the number and sophistication of systems. For example, TRIADS (Tripartite Assessment Delivery System) (Mackenzie 1999, TRIADS 2013), originated at the University of Derby and has been in constant development and use since 1992. TRIADS deliberately includes a wide variety of different question types to facilitate the testing of higher-order skills.

The STOMP (Software Teaching of Modular Physics) assessment system has been in development since

1995 (R.A. Bacon 2003). The latest version of STOMP (Bacon 2011) is a direct implementation of the QTI v2.1 specification, where the QTI (Question and Test Interoperability) specification is designed to enable the exchange of item, test and results data between authoring tools, item banks and assessment delivery systems etc. (IMS Global Learning Consortium 2013).

Also in the 1990s, there was growing concern about the mathematical preparedness of undergraduate students of physics and engineering, which led to the DIAGNOSYS diagnostic test (Appleby *et al.* 1997, Appleby 2007). DIAGNOSYS assumes a hierarchy of skills and employs an expert system to decide which question to ask next on the basis of a student's answers to previous questions. DIAGNOSYS is thus an early example of an adaptive test.

Around the turn of the century there was a move towards the online delivery of computer-marked assessment, using the internet to reach remote locations. At the UK Open University, interactive questions had been developed for a module that was first presented to students in 1997, using a precursor to the OpenMark system, but these questions were initially sent to students on a CD-ROM. It was not until 2002 that there was adequate confidence that students, studying in their own homes, would have sufficiently robust access to the internet to use the questions online (Jordan *et al.* 2003, Ross *et al.* 2006). Online delivery enables responses and scores to be recorded on servers at the Open University's headquarters in Milton Keynes. Reliable internet access also enables tutor-marked assignments to be submitted and returned electronically (Freake 2008, Jordan 2011), thus eliminating any delays caused by the postal system.

In the commercial sector, the company Questionmark (originally 'Question Mark') was founded in 1988. Question Mark for Web (launched in 1995) is believed to have been the world's first commercial web-based testing product. QuestionMark Professional was launched in 1993 and gradually superseded by Questionmark Perception (Kleeman 2013).

As more and more learning took place online, universities and other organisations started to use virtual learning environments (VLEs), also known as learning management systems. Most VLEs incorporate their own assessment systems. For example, the Moodle learning management system (Moodle 2013) was first released in 2002 and its quiz system has been in constant development since its first release in 2003 (Hunt 2012). Moodle and its assessment system are open source, reflecting a profound change in philosophy that

PART 2		ANSWER									
1	A	B	C	D	E	F	G	H	?	U	
2	A	B	C	D	E	F	G	H	?	U	
3	A	B	C	D	E	F	G	H	?	U	
4	A	B	C	D	E	F	G	H	?	U	
5	A	B	C	D	E	F	G	H	?	U	
6	A	B	C	D	E	F	G	H	?	U	
7	A	B	C	D	E	F	G	H	?	U	
8	A	B	C	D	E	F	G	H	?	U	
9	A	B	C	D	E	F	G	H	?	U	
10	A	B	C	D	E	F	G	H	?	U	

Figure 1 A machine-readable form used for entry of student responses to multiple-choice questions

has also influenced the development of e-assessment tools.

The current computer-marked assessment landscape

Selected response or constructed response?

Hunt (2012) identifies about around thirty different question types available within Moodle. The range of question types (e.g. 'drag-and-drop', 'calculated', 'numerical', 'true-false') adds variety, but Hunt's ad hoc survey of more than 50,000,000 questions from around 2,500 Moodle sites found that about 90% of the questions in use were selected-response questions, i.e. question types such as multiple-choice or drag-and-drop where options are presented for a student to select, in contrast to 'constructed-response' where students construct their own response.

The literature is littered with apparently contradictory evidence regarding the pros and cons of selected-response and constructed-response questions. Selected-response can assess a large breadth of knowledge (Betts *et al.* 2009, Ferrao 2010) whereas a test comprising constructed-response questions is likely to be more selective in its coverage. Use of selected-response questions also avoids issues of data-entry, particularly problematic in constructed-response questions when symbolic notation is required, for example in mathematics (Beevers & Paterson 2003, Jordan *et al.* 2003, Ross *et al.* 2006, Sangwin 2013). In addition, selected-response questions avoid issues with incomplete or inaccurate answer-matching. Occasional constructed-response answers may be incorrectly marked (Butcher & Jordan 2010). Gill & Greenhow (2008) report the worrying finding that students who had learned to omit units from their answers because these were not requested or could not be recognised by the assessment system, continued to omit units thereafter.

Conole & Warburton (2005) discuss the difficulty of using selected-response questions to assess higher order learning outcomes, though some have tried (e.g. Gwinnett & Cassella 2011). Furthermore, in some multiple-choice questions, the correct option can be selected by working back from the options, so the question is not assessing the learning outcome that it claims to be assessing. For example, a question that asks students to integrate a function can be answered by differentiating each of the options provided (Sangwin 2013). For all selected-response questions, especially those requiring a calculation or algebraic manipulation, if a student obtains an answer that is not one of the options

provided, they are given an early indication that there is likely to be something wrong with their answer (Bridgeman 1992). Even when testing the well-established force-concept inventory (first reported in Hestenes *et al.* 1992), Rebello & Zollman (2004) found that in equivalent open-ended tests, students gave answers that were not provided in the selected-response test.

Students may guess answers to selected-response questions, so their teacher has no way of telling what the student really understands (Crisp 2007). Downing (2003) is unconcerned about the impact of guessing on score, pointing out that it would be very difficult for a student to pass a whole assignment by guesswork alone. However, Burton (2005) points out that a successful guess has the potential to make a significant difference to the outcome for a borderline student.

Funk & Dickson (2011) used exactly the same questions in multiple-choice and short-answer free-text response format. Fifty students attempted both versions of each question, with half the students completing a 10 question short-answer pre-test before a 50 question multiple-choice exam and half the students completing the 10 short-answer questions as a post-test after the multiple-choice exam. In each case the performance on multiple-choice items was significantly higher ($p < 0.001$) than performance on the same items in the short-answer test. However, Ferrao (2010) found high correlation between scores on a multiple-choice and an open-ended test. Others have suggested that selected-response questions advantage particular groups of students, especially those who are more strategic or willing to take a risk (Hoffman 1967). Different gender biases have been reported, for example by Gipps & Murphy (1994) who found that 15-year old girls disliked multiple-choice questions whereas 15-year old boys preferred them to free-response types of assessment. Kuechler & Simkin (2003) found that students for whom English was a second language sometimes had difficulty dissecting the wording nuances of multiple-choice questions. Jordan & Mitchell (2009) and Nicol (2007) identify a fundamentally different cognitive process in answering selected-response and constructed-response questions.

Perhaps the most damning indictments of selected-response questions are those that query their authenticity. In commenting on the widespread use of multiple-choice questions in medical schools, Mitchell *et al.* (2003) quote Veloski *et al.* (1999): "Patients do not present with five choices". Bridgeman (1992, p271) makes a similar point with reference to engineers and chemists: they are seldom "confronted with five

numerical answers of which one, and only one, will be the correct solution”.

Back in the mid 1990s, Knight (1995, p13) pointed out that “what we choose to assess and how, shows quite starkly what we value”. Scouller (1998) argues that the use of selected-response questions can encourage students to take a surface approach to learning, although Kornell & Bjork (2007) found no support for the idea that students consider essay and short-answer tests to be more difficult than multiple-choice and so study harder for them. Roediger & Marsh (2005) and Marsh *et al.* (2007) found a diminished ‘testing effect’ when multiple-choice questions were used, attributed to the fact that the students were remembering the distractors rather than the correct answer.

The apparent contradictory results of investigations into the effectiveness of selected-response questions may be because the questions are not homogeneous (Simkin & Kuechler 2005). Different questions need different question types, with some questions (e.g. ‘Select the three equivalent expressions’) lending themselves particularly to a selected-response format. Burton (2005, p66) states that “It is likely that particular tests, and with them their formats and scoring methods, have sometimes been judged as unreliable simply because of flawed items and procedures.” Whatever question type is used, it is important that high-quality questions are written (Bull & McKenna 2000). For multiple-choice questions this means, for example, that all distractors should be equally plausible.

Even relatively simple multiple-choice questions can be used to create ‘moments of contingency’ (Dermo & Carpenter 2011) and Draper’s (2009) concept of catalytic assessment is based on the use of selected-response questions to trigger subsequent deep learning without direct teacher involvement. There are many ways in which the reliability and effectiveness of selected-response questions can be increased (Nicol 2007). Some of these techniques are discussed below.

Confidence-based marking and similar approaches

Various techniques have been used to compensate for the fact that students may guess the correct answers to multiple-choice questions. Simple negative marking (deducting marks or a percentage for incorrect answers) can be used, but care must be taken (Betts *et al.* 2009, Burton 2005).

Ventouras *et al.* (2010) constructed an examination using ‘paired’ multiple-choice questions on the same topic (but not obviously so to students), with a scoring rule which awarded a ‘bonus’ if they got both questions right. This gave results that were

statistically indistinguishable from the results of an examination with constructed-response questions. McAllister & Guidice (2012) describe another approach, in which the options are combined for a number of questions, resulting in a much longer list (60 options for 50 questions in their case) and so a much lower probability of guessing the correct answer. However, in general, it may be difficult to find options that are equally plausible for a range of questions.

Bush (2001) describes a ‘liberal multiple-choice test’ in which students may select more than one answer to a question if they are uncertain of the correct one. Negative marking is used to penalise incorrect responses: 3 marks are awarded for each correct solution, 1 mark is deducted for each incorrect solution and the total is divided by 3. If the student knows the right answer to the question, he or she can get $\frac{3}{3}$ i.e. 100% for that question. If a student is correct in thinking that the right answer is one of two options, he or she will get $\frac{(3-1)}{3}$ i.e. 67% for that question, rather than an equal chance of getting either 0% or 100%. If the student is correct in thinking that the right answer is one of three options, he or she will get $\frac{(3-2)}{3}$ i.e. 33% for that question, rather than having a 33% chance of getting 100% and a 67% chance of getting 0%. This approach is undoubtedly fairer, but some students found it confusing and concern has been expressed that it puts greater emphasis on tactics than on knowledge and understanding of the correct answer.

It has long been recognised (Ahlgren 1969) that the reliability of a test score can be increased by incorporating some sort of weighting for the appropriateness of a student’s confidence. Much work on ‘confidence-based’ (or ‘certainty-based’) marking has been done by Gardner-Medwin (2006), who notes that this approach does not favour the consistently confident or unconfident, but rather those who can correctly identify grounds for justification or reservation. Gardner-Medwin (2006) used the scale of marks and penalties shown in Table 1.

Rosewell (2011) required students to indicate their confidence *before* the multiple-choice options were revealed whilst Archer & Bates (2009) included a confidence indicator and also a free-text box into which students were required to give reasons for

Table 1 Marks and penalties for confidence-based marking (Gardner-Medwin 2006)

Confidence level	C=1 (low)	C=2 (mid)	C=3 (high)
Mark if correct	1	2	3
Penalty if incorrect	0	–2	–6

each answer. Nix & Wyllie (2011) incorporated both a confidence indicator and a reflective log into a formative multiple-choice quiz, in an attempt to encourage students to regulate their own learning experience.

‘Clickers’

Electronic voting systems, also known as ‘audience response systems’, ‘student response systems’ and ‘clickers’ have been used in classrooms and lecture theatres since before 1970. Students enter their answer to a multiple-choice question into a hand-held device of some sort and this relays information to their teacher or lecturer, who can then survey the understanding of the whole class and make appropriate adjustment to their teaching. Judson & Sawada’s (2002) historical review highlights the work of early pioneers like Boardman (1968) and Casanova (1971), who used a hard-wired system called an ‘Instructoscope’. The need for students to be able to enter their response in private was recognised from an early stage whilst Littauer (1972) provided the questions before class and noted students debating answers – an early indication of the type of approach later taken in Classtalk (Dufresne *et al.* 1996) and Peer Instruction (Mazur 1991). There is an extensive literature surrounding the use of clickers, with other reviews from Fies & Marshall (2006), Caldwell (2007), Simpson & Oliver (2007) and Kay & LeSage (2009). Online classrooms such as Blackboard Collaborate (Blackboard 2013) now enable similar voting to take place in a virtual environment.

Many authors (including Wieman 2010) attribute a profound positive effect on learning to the use of clickers, but Fies & Marshall (2006) call for more rigorous research whilst Beatty & Gerace (2009) argue that there are many different ways of using clickers and that these uses should not be lumped together. Peer discussion is found to be particularly effective, making a lecture more interactive and students more active participants in their own learning processes (Dufresne *et al.* 1996, Mazur 1991, Crouch & Mazur 2001, Lasry *et al.* 2008). ‘Dialogue around learning’ is one of Nicol & Macfarlane-Dick’s (2006) seven principles of good feedback practice and Nicol (2007) suggests that this can be achieved by initiating a class discussion of multiple-choice questions.

PeerWise

Nicol (2007) also points out that dialogue around learning can be achieved by having students work in small groups to construct multiple-choice questions or to comment on some aspect of tests that others have written. PeerWise (Denny *et al.* 2008b, PeerWise 2013) is a system developed in the

Computer Science Department at the University of Auckland but now in use worldwide, in which students author their own multiple-choice questions as well as using and evaluating questions written by their peers. Luxton-Reilly & Denny (2010) describe the pedagogy behind PeerWise, which rests on the premise that students shift from being consumers of knowledge to become participants in a community, producing and sharing knowledge. Evaluation at Auckland has shown that students consistently engage with the PeerWise system more than they are required to do (Denny *et al.* 2008c), that their questions are of remarkably high quality (Purchase *et al.* 2010) and that there is significant correlation between PeerWise activity and performance in subsequent written (not just multiple choice) questions (Denny *et al.* 2008a). These findings have been replicated in the School of Physics and Astronomy at the University of Edinburgh (Bates *et al.* 2012, Bates & Galloway 2013) and the correlation between PeerWise activity and subsequent performance was found to hold for the weaker students in the group as well as the stronger ones.

CALM, CUE and PASS-IT: focus on breaking a question down into ‘Steps’

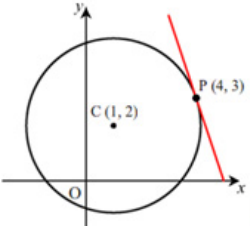
The CALM (Computer Aided Learning of Mathematics) Project started at Heriot-Watt University in 1985 (CALM 2001) and various computer-marked assessment systems have derived at least in part from CALM, including CUE, Interactive Past Papers, PASS-IT (Project on Assessment in Scotland – using Information Technology), i-assess and NUMBAS (Foster *et al.* 2012). Some of the systems have been used in high-stakes summative testing, but the focus has always been on supporting student learning (Ashton *et al.* 2006b). From the early days, constructed-response questions have been favoured, with hints provided to help students (Beevers & Paterson 2003).

One of the signatures of the CALM family of assessment systems is the use of ‘Steps’, allowing a question to be broken into manageable steps for the benefit of students who are not able to proceed without this additional scaffolding (Beevers & Paterson 2003, Ashton *et al.* 2006a). For the question shown in Figure 2a, a student could opt to work out the answer without intermediate assistance, and in summative use they would then be able to obtain full credit. Alternatively, they could click on ‘Steps’ at which point the question would be broken into separate steps as shown in Figure 2b. The student would then usually only be eligible for partial credit.

McGuire *et al.* (2002) compared the results for schoolchildren taking computer-marked tests in the CUE system with three different formats (no

Q5 Section A (40 marks) Exit 5 ⌂ ⌕

5.1) The diagram shows a circle, centre C (1, 2) and a tangent drawn at the point P (4, 3).



What is the equation of the tangent at P?

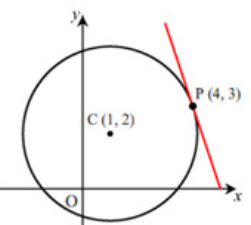
y =

Your submitted answer:

(a)

Q5 Section A (40 marks) Exit 5 ⌂ ⌕

5.1) The diagram shows a circle, centre C (1, 2) and a tangent drawn at the point P (4, 3).



What is the equation of the tangent at P?

y =

Your submitted answer: - 3x + 15

(b)

5.1.1) Find the gradient of the radial line CP and recall that the tangent and radius are perpendicular.

[0]

5.1.2) What is the gradient of the radial line CP?

Your submitted answer: $\frac{1}{3}$ ✓

[0.5] [0.5]

5.1.3) If m is the gradient of CP and n is the gradient of the tangent at P then

$m \times n = -1$

Your submitted answer: - 1 ✓

[0.5] [0.5]

5.1.4) What is the gradient of the tangent at P?

Your submitted answer: - 3 ✓

[0.5] [0.5]

5.1.1) Find the gradient of the radial line CP and recall that the tangent and radius are perpendicular.

[0]

5.1.2) What is the gradient of the radial line CP?

Your submitted answer: $\frac{1}{3}$ ✓

[0.5] [0.5]

5.1.3) If m is the gradient of CP and n is the gradient of the tangent at P then

$m \times n = -1$

Your submitted answer: - 1 ✓

[0.5] [0.5]

5.1.4) What is the gradient of the tangent at P?

Your submitted answer: - 3 ✓

[0.5] [0.5]

Figure 2 A question (a) before steps are revealed to the student; (b) after steps have been revealed. Reproduced by permission of the SCHOLAR Programme (Heriot-Watt University).

Steps, compulsory Steps or optional Steps) and with the partial credit they would have obtained by taking the corresponding examinations on paper. On this occasion no penalty was applied for the use of Steps. The overall marks for tests without Steps were lower than those in which Steps were available and they were also lower than the marks for the corresponding paper-based examinations. McGuire *et al.* concluded that “this means that without Steps the current marking schemes for paper-based examinations cannot, at present, be replicated by the current computer assessment packages. The longer and more sophisticated the question, the greater the problem.” They found no evidence of a difference in marks between what would be obtained from a paper-based examination or from a corresponding computer examination with Steps, whether optional or compulsory. However, they commented that even if the marks were similar “this does not mean that the candidates have shown the same skills. In particular, the use of Steps provides the candidate with the strategy to do a question.”

Ashton *et al.* (2004) describe PASS-IT’s use of summary reports on student performance. For students, the ability to see how they are performing

can be a driver towards independent learning. The analyses also feed into the cyclical question design process so, for example, if students consistently use STEPS in a particular question this may indicate that they are having difficulty starting this question. This may be due to a poorly designed question or it may reveal a student misunderstanding.

The research and software developments from PASS-IT are now fully integrated within the SCHOLAR Programme. SCHOLAR (2013), which deploys learning materials at Intermediate, Higher and Advanced Higher to all Scottish secondary schools, has formative assessment at its core.

OpenMark and Moodle: focus on feedback

The OpenMark system at the UK Open University was launched in 2005 following the success of interactive questions delivered to students by CD-ROM and the use of a precursor online system from 2002 (Jordan *et al.* 2003, Ross *et al.* 2006). The Open University’s large student numbers mean that investment in e-assessment is worthwhile; the fact that students are studying at a distance means that the provision of timely and targeted feedback is particularly important.

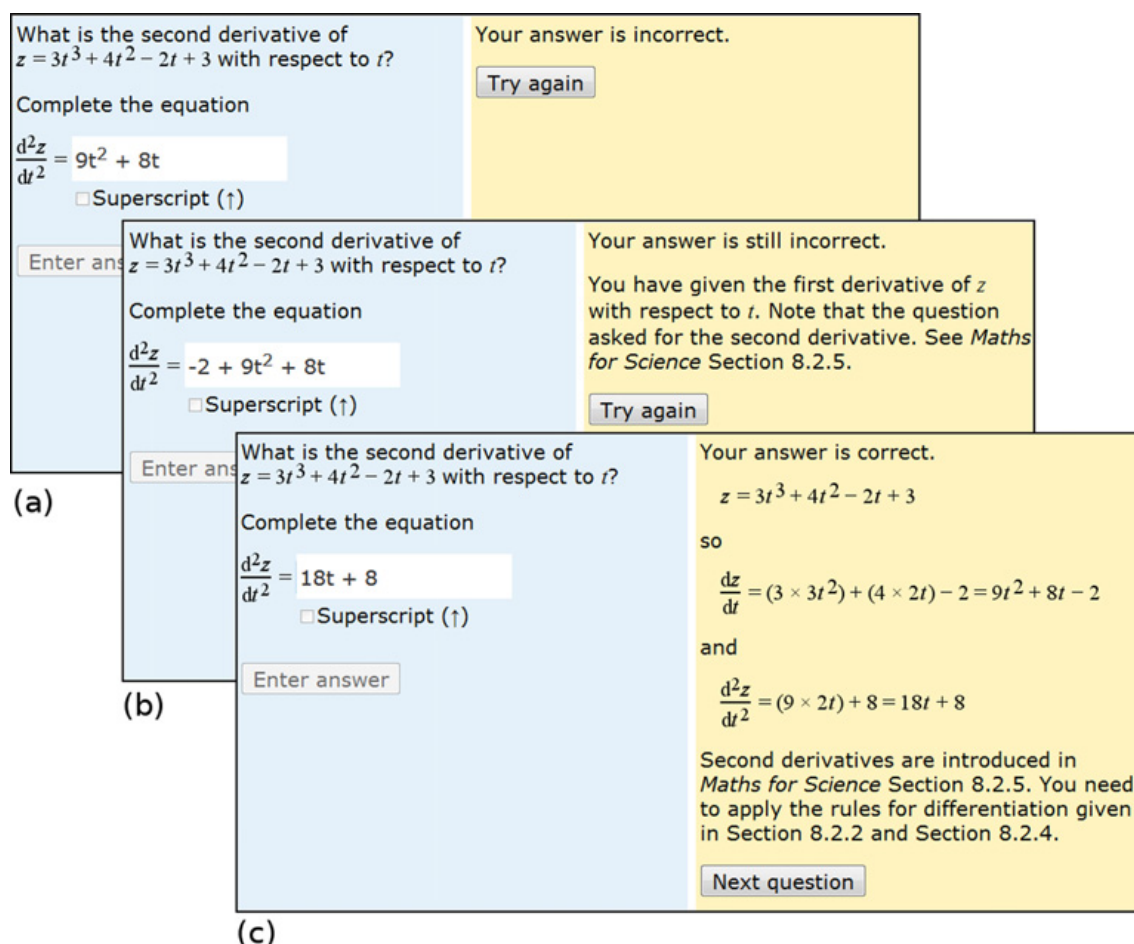


Figure 3 An OpenMark question showing feedback on repeated attempts at a question

A typical OpenMark question is shown in Figure 3, with the three screenshots representing a student's three attempts at the question. The general principles encapsulated in this example are:

- an emphasis on feedback;
- an emphasis on interactivity (so multiple attempts are provided with an opportunity for students so act immediately on the feedback received);
- the breadth of interactions supported (so a range of question types is available with the aim of "using the full capabilities of modern multimedia computers to create engaging assessments", Butcher 2008).

In addition, OpenMark assignments are designed to enable part-time students to complete them in their own time and in a manner that fits in with their everyday life. This means that they can be interrupted at any point and resumed later from the same location or from elsewhere on the internet (Butcher 2006, 2008).

Since 2002, interactive computer-marked assignments (iCMAs) have been introduced onto a range of Open University modules. In the year to August 2012, more than 630,000 iCMAs were

served for 60 separate modules (Butcher *et al.* 2013), with around a quarter of these being part of the module's formal assessment strategy (i.e. either summative or thresholded). In summative use, the credit awarded for a question reduces after each unsuccessful attempt. Appropriate credit and feedback can also be given for partially correct responses.

The Open University's Science Faculty was the first to embrace the use of OpenMark interactive computer-marked assignments (iCMAs) so the importance of answer-matching for correct units and precision in numerical answers was quickly realised (Ross *et al.* 2003). The need to provide appropriate targeted feedback on these points has been recognised as important so that a student appreciates the nature of their error. This feedback is usually given after a student's first attempt, even where most feedback is reserved for the second or third attempts. Jordan (2011 and 2012b) points out that altering the feedback provided on a question can sometimes have a significant impact on the way students react to the question. The use of statistical tools (Jordan *et al.* 2012) and the analysis of individual student responses (Jordan 2007) have led to improvements to the questions as well as giving insight into student misconceptions.

OpenMark's emphasis on the provision of a range of question types and on multiple tries with feedback influenced the development of Moodle's assessment system (Butcher 2008), and the Moodle authoring templates now allow for the provision of detailed targeted feedback for all Moodle question types, with a potential for the amount of feedback to be increased after successive student attempts. Hunt (2012) identifies question type (e.g. 'numerical' or 'drag-and-drop') and question behaviour (e.g. whether the question is to be run in 'interactive mode' with instantaneous feedback and multiple tries or in 'deferred mode' with just one attempt permitted and no feedback until the student's answers have been submitted) as separate concepts, and the combination of a question type and a question behaviour to generate an assessment item is a unique feature of Moodle.

Computer algebra-based systems (including STACK)

When free-text mathematical expressions are to be assessed, there are three ways in which a student's response can be checked. The original CALM assessment system evaluated the expression for particular numerical values. This is a reasonable approach, but is likely to lead to some incorrect responses being marked as correct (note, for example, that $2x$ and x^2 both have a value of 4 when $x = 2$). OpenMark uses string-matching. This has worked effectively (for example giving the targeted feedback shown in Figure 3 for the student answer given) but relies on the question-setter thinking of all

the equivalent answers that should be marked as correct and all the equivalent incorrect answers that should generate the same targeted feedback.

Since 1995, a number of computer-marked assessment systems have made use of a mainstream computer algebra system (CAS) to check student responses (Sangwin 2013). For example AIM uses the computer algebra system Maple (Strickland 2002), CABLE uses Axiom (Naismith & Sangwin 2004), whilst STACK (System for Teaching and Assessment using a Computer Algebra Kernel) chose the open source computer algebra system Maxima (Sangwin 2013).

STACK was first released in 2004 as a stand-alone system but since 2012 it has been available as a Moodle question type (Butcher *et al.* 2013). Aspects considered important by the Moodle system developers, in particular the provision of feedback and the monitoring of student answers, are also important in STACK. Behind the scenes, STACK employs a 'potential response tree' to enable tailored feedback to be provided in response to common errors. A focus group at Aalto University, Finland considered the immediate feedback to be the best feature of STACK (Sangwin 2013) and Sangwin (p104) expresses his surprise that "not all systems which make use of a CAS enable the teacher to encode feedback".

The question shown in Figure 4 illustrates some of STACK's sophisticated features, described in more detail by Sangwin:

Find $h'(x)$ where $h(x) = (x^5 + 3) \sin(3x)$.

$h'(x) = 3*(x^5+3)*\cos(3*x) + 5*x^4*\sin(3*x)$

Your last answer was interpreted as follows:

$$3 (x^5 + 3) \cos(3x) + 5x^4 \sin(3x)$$

Check

Correct answer, well done.

$h(x)$ can be expressed as a product

$$h(x) = f(x)g(x)$$

where $f(x) = x^5 + 3$ and $g(x) = \sin(3x)$, so one way to differentiate $h(x)$ is to use the Product Rule.

$$f'(x) = 5x^4$$

$$g'(x) = 3 \cos(3x)$$

Therefore

$$h'(x) = f'(x)g(x) + f(x)g'(x)$$

$$= (5x^4)(\sin(3x)) + (x^5 + 3)(3 \cos(3x))$$

This answer can be expressed in several different forms.

See Unit 1 Section 5.2 Differentiating combinations of functions.

Figure 4 A correctly answered STACK question

Variants of questions. Since all calculations are performed by the computer algebra system, variables can be assigned in such a way that many different variants of a question can be authored for minimal effort. So, in the example shown, variants might require students to differentiate any function that is a product of two other functions. However, best practice is to check and deploy only selected variants of similar difficulty.

Author choice. The question author can decide, for example, whether to accept implicit multiplication and whether to insert a dot or cross to indicate where multiplication has been assumed. The author can also choose whether to accept all algebraically equivalent answers. This has been done in the example shown in Figure 4, but in answer to the question 'Simplify u^3u^5 ', an answer of ' u^3u^5 ' would not be acceptable.

Validation is separated from assessment. The answer that has been input by the student using keyboard functions has been displayed as the system has 'understood' it, and its syntactic validity has been checked before marking. This also gives the opportunity for answers that are trying to trick the computer algebra system, in this case answers containing the command 'Diff', which tells the computer algebra system to differentiate the original function, to be rejected.

Short answer questions and essays

Alongside the use of computer algebra systems for more sophisticated mathematical questions, the introduction of software for marking short-answer questions has extended the range of constructed response questions that can be used. 'Short-answer' is usually taken to mean questions requiring

answers of a sentence or two in length and, following evaluation, Jordan (2012b) restricts the length to no more than 20 words, partly to give an indication to students of what is required and partly to discourage responses that include both correct and incorrect aspects. Mitchell *et al.* (2002) first recognised the incorrect qualification of a correct answer as a potentially serious problem for the automatic marking of short-answer questions.

Software for marking short-answer questions includes C-rater (Leacock & Chodorow 2003) and systems developed by Intelligent Assessment Technologies (IAT) (Mitchell *et al.* 2002, Jordan & Mitchell 2009) and by Sukkarieh *et al.* (2003, 2004). These systems, reviewed by Siddiqi & Harrison (2008), are all based to some extent on computational linguistics. For example, the IAT software draws on the natural language processing (NLP) techniques of information extraction and compares templates based around the verb and subject of model answers with each student response. However, IAT provide an authoring tool that can be used by a question author with no knowledge of NLP. In contrast OpenMark's PMatch (Butcher & Jordan 2010, Jordan 2012a) and the Moodle Pattern Match question type are simpler pattern matching systems, based on the matching of keywords and their synonyms, sometimes in a particular order and/or separated by no more than a certain number of other words, and with consideration paid to the presence or absence of negation (as shown in Figure 5). A dictionary-based spell checker notifies students if their response contains a word that is not recognised, but the standard string-matching techniques of allowing missing or transposed letters remains useful, to cope with situations where a student accidentally

A hailstone falls vertically with a constant speed. What does this tell you about the forces acting on the hailstone?

Please give your answer as a **short phrase or sentence.**

This tells me that there are no unbalanced forces acting on the hailstone.

Enter answer

Your answer is correct.

Since the hailstone is falling with constant velocity the forces on it must be balanced. This is a consequence of Newton's First Law of Motion.

air resistance

weight

Next question

Figure 5 A correctly answered PMatch question

uses a word that is slightly different from the intended one (e.g. 'decease' instead of 'decrease'). For most of the systems, the fact that real student responses are used in developing answer-matching is regarded as being of crucial significance.

IAT and PMatch answer matching at the Open University (Jordan & Mitchell 2009, Butcher & Jordan 2010) was used within OpenMark whilst Pattern Match is a Moodle Question type so, as with the STACK question type, instantaneous and tailored feedback is considered very important. Jordan (2012b) has conducted a detailed evaluation of student engagement with short-answer free-text questions and the feedback provided.

Good marking accuracy has been obtained, always comparable or better than the marking of human markers (Butcher & Jordan 2010, Jordan 2012a), but yet questions of this type remain under-used. Jordan (2012a) identifies the need to collect and mark several hundred student responses and the time taken to develop answer matching as significant barriers to wider uptake. She suggests that research should be focused on the use of machine learning for the development of answer-matching rules and on an investigation into the extent to which students from different universities give the same answers to questions of this type; if their answers are similar then there is the potential to share questions.

Automated systems for the marking of essays are characteristically different from those used to mark short-answer questions, because with essay-marking systems the focus is frequently on the writing style, and the required content can be less tightly constrained than is the case for shorter answers. Many systems exist for the automatic marking of essays, for example E-rater (Attali & Burstein 2006) and Intelligent Essay Assessor (Landauer 2003), with reviews by Valenti *et al.* (2003), Dikli (2006) and Vojak *et al.* (2011). Further systems are under development and some, for example OpenEssayist (Van Labeke *et al.* 2013), put the focus on the provision of feedback to help students to develop their essay-writing skills. Systems that use simple proxies for writing style have been criticised, for example by Perelman (2008) who trained three students to obtain good marks for a computer-marked essay by such tricks as using long words and including a famous quotation (however irrelevant) in the essay's conclusion. Condon (2013) contends that until computers can make a meaningful assessment of writing style, they should not be used.

Using questions effectively

According to Hunt (2012) a computer-marked assessment system comprises three parts:

- a question engine that presents each question to the student, grades their response and delivers appropriate feedback on that question;
- a question bank;
- a test system that combines individual items into a complete test (possibly with feedback at the test level).

Thus, in addition to considering question types, it is necessary to consider the way in which they are combined.

The selection of questions from a question bank or the use of multiple variants of each question can provide additional opportunities for practice (Sangwin 2013) and discourage plagiarism (Jordan 2011). However, especially in summative use, it is necessary to select questions that assess the same learning outcome and are of equivalent difficulty (Dermo 2010, Jordan *et al.* 2012). Dermo (2009) found a concern among students that the random selection of items was unfair.

Feedback can be given to students at the level of the quiz or test. For example, Jordan (2011) describes a diagnostic quiz in which a traffic light system is used to indicate students' preparedness in a number of different skill areas, with 'red' meaning 'you are not ready', 'green' meaning 'you appear to have the requisite skills' and 'amber' meaning 'take care'.

Adaptive assessments (frequently described as 'computer adaptive tests') use a student's responses to previous questions to make a judgement about his or her ability, and so to present subsequent questions that are deemed to be at an appropriate level (Crisp 2007). Lilley *et al.* (2004) found that students were not disadvantaged by a computer adaptive test and that they appreciated not having to answer questions that they considered too simple. Questions for computer adaptive tests are usually selected from a question bank and statistical tools are used to assign levels of difficulty (Gershon 2005), thus most systems become complicated and rely on large calibrated question banks. Pyper & Lilley (2010) describe a simpler 'flexilevel' system which applies fixed branching techniques to select the next item to present to a student at each stage.

Another use of adaptive testing is to create a 'maze' in which questions are asked that depend on a student's answer to the previous question, without necessarily attributing 'correctness' or otherwise. Wyllie & Waights (2010) developed a clinical decision-making maze to simulate the decisions that have to be taken, based on various sources of information, in deciding how to treat an elderly patient with a leg ulcer. This type of maze offers

one way in which the authenticity of computer-marked assessments might be increased.

The CALM Team at Heriot-Watt University has sought to integrate assessment with simulations (Ashton & Thomas 2006), for example using a split screen with simulated practical work on one side and a question on the other (Thomas & Milligan 2003), with the aim of providing rich interactivity and assessing students “as they learn, in the environment in which they learn” (Ashton *et al.* 2006b, p125).

In another attempt to align teaching and assessment, many textbooks are accompanied by question banks, with one of the best known products being Pearson’s ‘Mastering’ series, such as MasteringPhysics® (Pearson 2013). Despite the fact that such questions are often described as ‘homework’, teachers retain the choice of how to employ them, and Pearson’s tools for analysing student responses have led the way in their provision of information about student misunderstandings. For example, Walet & Birch (2012) use a model of ‘just in time’ teaching in which a lecture sets the scene then students do self-study prior to an online quiz. The results of the quiz inform the content of classes a few hours later. Walet & Birch use MasteringPhysics® but found that questions in ‘Homework mode’ (with no tips and hints) were not well received by their students, so they now use ‘Tutorial mode’ (with tips and hints) and have added feedback where necessary.

E-assessment beyond quizzes

Technology can be used to support assessment and the delivery of feedback in myriad other ways. Assignments can be submitted, human-marked and returned online; Hepplestone *et al.* (2011) review work in this area and describe a system which releases feedback to students but stalls the release of grades to them until they have reflected on the feedback received. This approach significantly enhanced students’ engagement with the feedback (Parkin *et al.* 2012). Similarly positive results have been reported for the use of audio feedback (Lunt & Curran 2010, McGarvey & Haxton 2011) and screencasting (Haxton & McGarvey 2011, O’Malley 2011).

The use of PeerWise for student authoring and review of questions was discussed earlier, but peer assessment more generally refers to the assessing of students’ work by their peers. This can give students additional feedback for minimum staff resource, but Honeychurch *et al.* (2013) point out that the real value of peer assessment “resides not in the feedback (the product) but in the process of creating the feedback”. Technology can be used to

support peer assessment, making it available to large class sizes and in online environments (Luxton-Reilly 2009, Honeychurch *et al.* 2013).

Technologies such as e-portfolios, blogs, wikis and forums can be used to encourage student engagement, collaboration and reflection (Bennett *et al.* 2012). E-portfolios (‘electronic portfolios’) such as Pebblepad (2013) are widely used to enable students to record and reflect on their learning and to store evidence of achievement, usually across a range of modules. This enables the assessment of skills that are otherwise difficult to assess and encourages a reflective approach to learning, with students having responsibility for what they include and so being able to focus on the positive (Jafari & Kaufman 2006, Madden 2007). Teachers can access the e-portfolios of their students for the purposes of assessment and feedback and Molyneaux *et al.* (2009) describe a project in which an e-portfolio was jointly managed by a group of students. Portfolios are frequently associated with the enhancement of student employability (Halstead & Sutherland 2006).

Sim & Hew (2010) recognise the potential of blogs as a natural partner to e-portfolios, enabling students to share their experiences and reflection. Churchill (2009) encouraged students to engage in a blogging activity by requiring participation as part of the course assessment. The activity was well received, with just over half of the students reporting that they would continue with the blog even if not assessed.

Caple & Bogle (2013) are enthusiastic about the usefulness of wikis in assessing collaborative projects. Wiki pages can be modified by any member of the group, but their particular advantage for assessment is that each modification is recorded and attributed to a specific user. This means that the work of the group can be assessed but also the work of each individual. Online forums are also useful tools for collaboration, but Conole & Warburton (2005) recognise the difficulty in ‘measuring’ different interactions on a forum. Software tools may be of assistance (e.g. Shaul 2007), though human markers will still be required to assess the quality of a student’s contribution.

E-assessment: what is the future?

On the basis of the e-assessment literature reviewed here and related educational and technological developments, it is possible to make predictions for the future of e-assessment and to discuss potential pitfalls.

Massive online learning

In summer 2013, Cathy Sandeen wrote in a special edition of the *Journal of Research & Practice in Assessment* that “MOOCs [Massive Open Online Courses] have focused our attention and have fostered much excitement, experimentation, discussion and debate like nothing I have seen in higher education” (Sandeen 2013, p11). Whatever the future of MOOCs, a beneficial side effect is that they are forcing the assessment community to consider appropriate methodologies for assessing huge student numbers and for assessing informal and online learning for the future.

MOOCs were originally offered at no cost to students and on a no credit basis. At this level, Masters’ (2011) assertion that “In a MOOC, assessment does not drive learning; learners’ own goals drive learning” is reasonable. However, most MOOCs issue ‘badges’ of some sort, based on a certain level of engagement or attainment and if attainment is to be measured then assessment of some sort is required. Additionally, formative assessment offers the possibility to engage and motivate students and to provide them with feedback, all of which factors might act to improve the currently abysmal completion rates of most MOOCs.

Learning from MOOCs can be most easily assessed by computer-marked assessment and collaborative and peer assessment operating in an entirely online environment. Questions for computer-marked assessment need to be delivered at low cost and quickly and there is a danger that this will lead to poor quality assessment. To avoid this, MOOC systems need to provide a variety of question types, with the potential for instantaneous meaningful feedback and for students to attempt the question several times. It is also important that different students receive a different set of questions, so question banks or multiple variants of questions are required. Finally, after providing high quality tools, attention should be paid to ensuring that MOOC authors are trained to write high quality questions. Teachers should not be “neglected learners” (Sangwin & Grove 2006).

Learning analytics and assessment analytics

Learning analytics can be defined as “the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs” (Ferguson 2012, p305) and Redecker *et al.* (2012) suggest that we should “move beyond the testing paradigm” and start employing learning analytics in assessment. Data collected from student interaction in an online

environment offers the possibility to assess students on their actual interactions rather than adding assessment as a separate event. Clow (2012) points out that learning analytics systems can provide assessment-like feedback even in informal settings.

At a broader level, learning analytics can inform teachers about the learning of a cohort of students and Ellis (2013) calls for ‘assessment analytics’ (the analysis of assessment data), pointing out that assessment is ubiquitous in higher education whilst student interactions in other online learning environments (in particular social media) are not. The potential for work of this type is illustrated by Jordan (2013), who has analysed student behaviour on computer-marked assignments, uncovering specific student misunderstandings but also revealing more about the drivers for deep engagement. She also found significantly different patterns of use by two cohorts of students, on assignments known to be equivalent, and was able to link the different patterns of use to differences between the two populations of students and their workload.

Blurring of the boundaries between teaching, assessment and learning

If, as suggested by the papers reviewed earlier, we use learning analytics as assessment and we use assessment analytics to learn more about learning, the boundaries between assessment and learning become blurred. Similarly, more sophisticated expert systems should be able to deliver better adaptive tests, offering the potential for formative and diagnostic (and perhaps summative) assessment that is personalised to the level of each student, and so better able to support their learning.

The online delivery of teaching, perhaps on mobile devices, enables smooth progression from teaching resources to assessment and back again. Questions can be asked at the most appropriate point in the teaching, and attempted wherever and whenever the student is doing their studying. An assignment will not always be a separate entity. This is a familiar concept from in-text questions in textbooks, but now the questions can be interactive. A range of teaching resources can easily be accessed to help the student answer the question, but any ‘model answer’ can remain hidden until they have submitted their response.

Appropriate but not inappropriate use of a computer

Speaking at the eSTeEM (2013) Annual Conference in 2013, Phil Butcher reviewed the use of computer-marked assessment at the Open University and then said “Now what? Might I suggest starting to

use the computer as a computer?" He was referring specifically to the introduction of a STACK question type into Moodle (Butcher *et al.* 2013). However, there are other recent examples of work being explicitly and authentically marked by a computer, for example 'Coderunner' (Lobb 2013), used to assess computer programming skills by actually running the code that the student has written. In addition to using a computer to compute and evaluate in the assessment itself, it should be possible to harness technology to improve the quality of our questions, for example by using machine learning to develop answer-matching rules from marked student responses to short-answer free-text questions (Jordan 2012a).

However, computers should be used only when it is appropriate to do so and sometimes a hybrid approach is more effective. Sail-M (Semi-automatic analysis of individual Learning processes in Mathematics) automatically monitors student interactions and then, if necessary, passes these to a tutor who supplied detailed feedback (Herding & Schroeder 2012). Butcher & Jordan (2010) suggest that short-answer responses that the computer does not 'recognise' might be passed to a human marker. At present, there remain some

assessed tasks (e.g. experimental reports, essays, proofs) that present considerable challenges for machine marking. It is reasonable to use essay-marking software for formative purposes, but questions have been raised about its widespread summative use and, as recognised by McGuire *et al.* (2002) more than 10 years ago, we must not fool ourselves that systems that break problems into steps are assessing all the skills involved in problem solving. In summary, we should use computers to do what they do best, relieving human markers of some of the drudgery of marking and freeing up time for them to assess what they and only they can assess with authenticity.

Acknowledgements

The author gratefully acknowledges inspiration and assistance from Tom Mitchell of Assessment Technologies Ltd, Chris Sangwin at the University of Birmingham (developer of the STACK system), Cliff Beevers, Emeritus Professor at Heriot-Watt University, Oormi Khaled, Educational Developer (Assessment), Heriot-Watt University, and, in particular, from Phil Butcher and Tim Hunt at the Open University.

References

- Ahlgren, A. (1969) Reliability, predictive validity, and personality bias of confidence-weighted scores. In *Symposium on Confidence on Achievement Tests: Theory, Applications, at the 1969 joint meeting of the American Educational Research Association and the National Council on Measurement in Education*.
- Angus, S.D. and Watson, J. (2009) Does regular online testing enhance student learning in the numerical sciences? Robust evidence from a large data set. *British Journal of Educational Technology* **40** (2), 255–272.
- Appleby, J. (2007) DIAGNOSYS. Available at <http://www.staff.ncl.ac.uk/john.appleby/diagpage/diagindx.htm> (accessed 7 June 2013).
- Appleby, J., Samuels, P. and Treasure-Jones, T. (1997) DIAGNOSYS: A knowledge-based diagnostic test of basic mathematical skills. *Computers & Education* **28** (2), 113–131.
- Archer, R. and Bates, S. (2009) Asking the right questions: Developing diagnostic tests in undergraduate physics. *New Directions in the Teaching of Physical Sciences* **5**, 22–25.
- Ashburn, R. (1938) An experiment in the essay-type question. *Journal of Experimental Education* **7** (1), 1–3.
- Ashton, H.S., Beevers, C.E., Schofield, D.K. and Youngson, M.A. (2004) Informative reports: Experience from the PASS-IT project. In *Proceedings of the 8th International Computer Assisted Assessment Conference, Loughborough*.
- Ashton, H.S., Beevers, C.E., Korabinski, A.A. and Youngson, M.A. (2006a) Incorporating partial credit in computer-aided assessment of Mathematics in secondary education. *British Journal of Educational Technology* **37** (1), 93–119.
- Ashton, H.S., Beevers, C.E., Milligan, C.D., Schofield, D.K., Thomas, R.C. and Youngson, M.A. (2006b) Moving beyond objective testing in online assessment. In *Online assessment and measurement: Case studies from higher education, K-12 and corporate* (ed. S.C. Howell and M. Hricko), pp116–127. Hershey, PA: Information Science Publishing.
- Ashton, H.S. and Thomas, R.C. (2006) Bridging the gap between assessment, learning and teaching. In *Proceedings of the 10th International Computer Assisted Assessment Conference, Loughborough*.

- Attali, Y. and Burstein, J. (2006) Automated essay scoring with E-rater® V.2. *The Journal of Technology, Learning & Assessment*, **4** (3).
- Bacon, D.R. (2003) Assessing learning outcomes: A comparison of multiple-choice and short answer questions in a marketing context. *Journal of Marketing Education* **25** (1), 31–36.
- Bacon, R.A. (2003) Assessing the use of a new QTI assessment tool within Physics. In *Proceedings of the 7th International Computer Assisted Assessment Conference, Loughborough*.
- Bacon, R.A. (2011) *Software Teaching of Modular Physics*. Available at <http://www.stomp.ac.uk/> (accessed 7 June 2013).
- Bates, S. and Galloway, R.K. (2013) *Student generated assessment*. Education in Chemistry, January 2013.
- Bates, S.P., Galloway, R.K. and McBride, K.L. (2012) Student-generated content: Using PeerWise to enhance engagement and outcomes in introductory physics courses. In *Proceedings of the 2011 Physics Education Research Conference, Omaha, Nebraska*.
- Beatty, I. and Gerace, W. (2009) Technology-enhanced formative assessment: a research-based pedagogy for teaching science with classroom response technology. *Journal of Science Education and Technology* **18** (2), 146–162.
- Beevers, C. and Paterson, J. (2003) Automatic assessment of problem solving skills in mathematics. *Active Learning in Higher Education* **4** (2), 127–144.
- Beevers, C. et al. (2010) What can e-assessment do for learning and teaching? Part 1 of a draft of current and emerging practice: review by the E-Assessment Association Expert Panel (presented by John Winkley of AlphaPlus on behalf of the panel). In *Proceedings of CAA 2010 International Conference, Southampton* (ed. D. Whitelock, W. Warburton, G. Wills and L. Gilbert).
- Bennett, S., Bishop, A., Dalgarno, B., Waycott, J. and Kennedy, G. (2012) Implementing Web 2.0 technologies in higher education: a collective case study. *Computers & Education* **59** (2), 524–534.
- Betts, L.R., Elder, T.J., Hartley, J. and Trueman, M. (2009) Does correction for guessing reduce students' performance on multiple-choice examinations? Yes? No? Sometimes? *Assessment & Evaluation in Higher Education* **34** (1), 1–15.
- Blackboard (2013) *Blackboard Collaborate*. Available at <http://www.blackboard.com/Platforms/Collaborate/Overview.aspx> (accessed 7 June 2013).
- Boardman, D.E. (1968) *The use of immediate response systems in junior college*. Unpublished Master's thesis, University of California, Los Angeles. Eric Document Reproduction Service ED019057.
- Bridgeman, B. (1992) A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement* **29**, 253–271.
- Brown, S., Bull, J. and Race, P. (1999) *Computer-assisted assessment in higher education*. London: Kogan Page.
- Bull, J. and Danson, M. (2004) *Computer-aided assessment (CAA)*. York: LTSN Generic Centre.
- Bull, J. and McKenna, C. (2000) Quality assurance of computer-aided assessment: Practical and strategic issues. *Quality Assurance in Education* **8** (1), 24–31.
- Bull, J. and McKenna, C. (2004) *Blueprint for computer-aided assessment*. London: Routledge Falmer.
- Burton, R.F. (2005) Multiple-choice and true/false tests: myths and misapprehensions. *Assessment & Evaluation in Higher Education* **30** (1), 65–72.
- Bush, M. (2001) A multiple choice test that rewards partial knowledge. *Journal of Further & Higher Education* **25**, 157–163.
- Butcher, P.G. (2006) *OpenMark Examples*. Available at <http://www.open.ac.uk/openmarkexamples> (accessed 7 June 2013).
- Butcher, P.G. (2008) Online assessment at the Open University using open source software: Moodle, OpenMark and more. In *Proceedings of the 12th International Computer Assisted Assessment Conference, Loughborough*.
- Butcher, P.G., Hunt, T.J. and Sangwin, C.J. (2013) Embedding and enhancing e-Assessment in the leading open source VLE. In *Proceedings of the HEA-STEM Annual Conference, Birmingham*.
- Butcher, P.G. and Jordan, S.E. (2010) A comparison of human and computer marking of short free-text student responses. *Computers & Education* **55**, 489–499.
- Caldwell, J.E. (2007) Clickers in the large classroom: current research and best-practice tips. *Life Sciences Education* **6**, 9–20.
- CALM (2001) *Computer Aided Learning in Mathematics*. Available at <http://www.calm.hw.ac.uk/> (accessed 7 June 2013).
- Caple, H. and Bogle, M. (2013) Making Group assessment transparent: what wikis can contribute to collaborative projects. *Assessment & Evaluation in Higher Education* **38** (2), 198–210.
- Casanova, J. (1971) An instructional experiment in organic chemistry: the use of a student response

- system. *Journal of Chemical Education* **48** (7), 453–455.
- Churchill, D. (2009) Educational applications of Web 2.0: Using blogs to support teaching and learning. *British Journal of Educational Technology* **40** (1), 179–183.
- Clow, D. (2012) The learning analytics cycle: closing the loop effectively. In *Proceedings of the 2nd International Conference on Learning Analytics & Knowledge, Vancouver, BC*.
- Condon, W. (2013) Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing* **18**, 100–108.
- Conole, G. and Warburton, B. (2005) A review of computer-assisted assessment. *Research in Learning Technology* **13** (1), 17–31.
- Crisp, G. (2007) *The e-Assessment Handbook*. London, Continuum.
- Crouch, C. H. and Mazur, E. (2001) Peer instruction: ten years of experience and results. *American Journal of Physics* **69** (9), 970–977.
- Denny, P., Hamer, J., Luxton-Reilly, A. and Purchase, H. (2008a) PeerWise: students sharing their multiple choice questions. In *Proceedings of the Fourth international Workshop on Computing Education Research*. pp51–58.
- Denny, P., Luxton-Reilly, A. and Hamer, J. (2008b) The PeerWise system of student contributed assessment questions. In *Proceedings of the 10th Conference on Australasian Computing Education*, pp69–74.
- Denny, P., Luxton-Reilly, A. and Hamer, J. (2008c) Student use of the PeerWise system. *ACM SIGCSE Bulletin* **40** (3), 73–77.
- Dermo, J. (2007) Benefits and obstacles: factors affecting the uptake of CAA in undergraduate courses. In *Proceedings of the 11th International Computer Assisted Assessment Conference, Loughborough*.
- Dermo, J. (2009) e-Assessment and the student learning experience: a survey of student perceptions of e-assessment. *British Journal of Educational Technology* **40** (2), 203–214.
- Dermo, J. (2010) In search of Osiris: Random item selection, fairness and defensibility in high-stakes e-assessment. In *Proceedings of CAA 2010 International Conference, Southampton* (eds. D. Whitelock, W. Warburton, G. Wills & L. Gilbert).
- Dermo, J. and Carpenter, L. (2011) e-Assessment for learning: can online selected response questions really provide useful formative feedback? In *Proceedings of CAA 2011 International Conference, Southampton* (eds. D. Whitelock, W. Warburton, G. Wills and L. Gilbert).
- Dikli, S. (2006) An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment* **5** (1).
- Downing, S.M. (2003) Guessing on selected-response examinations, *Medical Education* **37** (8), 670–671.
- Draper, S.W. (2009) Catalytic assessment: understanding how MCQs and EVS can foster deep learning. *British Journal of Educational Technology* **40** (2), 285–293.
- Dufresne, R.J., Wenk, L., Mestre, J.P., Gerace, W.J. and Leonard, W.J. (1996) Classtalk: A classroom communication system for active learning. *Journal of Computing in Higher Education* **7**, 3–47.
- Earley, P.C. (1988) Computer-generator performance feedback in the subscription processing industry. *Organizational Behavior and Human Decision Processes* **41**, 50–64.
- eSTEEeM (2013) Available at <http://www.open.ac.uk/about/teaching-and-learning/esteem/> (Accessed 7 June 2013).
- Ellis, C. (2013) Broadening the scope and increasing the usefulness of learning analytics: The case for assessment analytics. *British Journal of Educational Technology*, **44** (4), 662–664.
- Ferguson, R. (2012) Learning analytics: drivers, developments and challenges, *International Journal of Technology Enhanced Learning* **4** (5/6), 304–317.
- Ferrao, M. (2010) E-assessment within the Bologna paradigm: evidence from Portugal. *Assessment & Evaluation in Higher Education* **35** (7), 819–830.
- Fies, C. and Marshall, J. (2006) Classroom response systems: A review of the literature. *Journal of Science Education and Technology* **15** (1), 101–109.
- Foster, B., Perfect, C. and Youd, A. (2012) A completely client-side approach to e-assessment and e-learning of mathematics and statistics. *International Journal of e-Assessment* **2** (2).
- Freake, S. (2008) Electronic marking of physics assignments using a tablet PC. *New Directions in the Teaching of Physical Sciences* **4**, 12–16.
- Funk, S.C. and Dickson, K.L. (2011) Multiple-choice and short-answer exam performance in a college classroom. *Teaching of Psychology* **38** (4), 273–277.
- Gardner-Medwin, A.R. (2006). Confidence-based marking: towards deeper learning and better exams. In *Innovative Assessment in Higher Education* (ed. C. Bryan and K. Clegg). London: Routledge. pp141–149.

- Gershon, R.C. (2005) Computer adaptive testing. *Journal of Applied Measurement* **6** (1), 109–127.
- Gill, M. and Greenhow, M. (2008) How effective is feedback in computer-aided assessments? *Learning Media & Technology* **33** (3), 207–220.
- Gipps, C.V. (2005) What is the role for ICT-based assessment in universities? *Studies in Higher Education* **30** (2), 171–180.
- Gipps, C. and Murphy, P. (1994) *A fair test? Assessment, achievement and equity*. Buckingham: Open University Press.
- Grebenik, P. and Rust, C. (2002) IT to the rescue. In *Assessment: case studies, experience and practice in Higher Education* (eds. P. Schwartz and G. Webb). London: Kogan Page.
- Gwinnett, C. and Cassella, J. (2011) The trials and tribulations of designing and utilising MCQs in HE and for assessing forensic practitioner competency, *New Directions in the Teaching of Physical Sciences* **7**, 72–78.
- Halstead, A. and Sutherland, S. (2006) ePortfolio: A means of enhancing employability and the professional development of engineers. In *International Conference on Innovation, Good Practice and Research in Engineering Education, Liverpool*.
- Haxton, K.J. and McGarvey, D.J. (2011) Screencasts as a means of providing timely, general feedback on assessment. *New Directions in the Teaching of Physical Sciences* **7**, 18–21.
- Hepplestone, S., Holden, G., Irwin, B., Parkin, H.J. and Thorpe, L. (2011) Using technology to encourage student engagement with feedback: a literature review. *Research in Learning Technology* **19** (1), 117–127.
- Herding, D. and Schroeder, U. (2012) Using capture and replay for semi-automatic assessment, *International Journal of e-Assessment* **2** (1).
- Hestenes, D., Wells, M. and Swackhamer, G. (1992) Force concept inventory. *The Physics Teacher* **30**, 141–158.
- Hoffman, B. (1967) Multiple-choice tests. *Physics Education* **2**, 247–51.
- Honeychurch, S., Barr, N., Brown, C. and Hamer, J. (2013) Peer assessment assisted by technology. *International Journal of e-Assessment* **3** (1).
- Hunt, T.J. (2012) Computer-marked assessment in Moodle: Past, present and future. In *Proceedings of CAA 2012 International Conference, Southampton* (ed. D. Whitelock, W. Warburton, G. Wills and L. Gilbert).
- IMS Global Learning Consortium (2013) IMS question and test interoperability specification. Available at <http://www.imsglobal.org/question/> (accessed 7 June 2013).
- Jafari, A. and Kaufman, C. (2006) *Handbook of Research on ePortfolios*. Hershey, PA: Idea Group Publishing.
- JISC (2006) *e-Assessment Glossary (Extended)*. Available at http://www.jisc.ac.uk/uploaded_documents/eAssess-Glossary-Extended-v1-01.pdf (accessed 7 June 2013).
- JISC (2009) *Review of Advanced e-Assessment Techniques (RAeAT)*. Available at <http://www.jisc.ac.uk/whatwedo/projects/raeat> (accessed 7 June 2013).
- JISC (2010) *Effective assessment in a digital age : a guide to technology-enhanced assessment and feedback*. Available at <http://www.jisc.ac.uk/publications/programmerelated/2010/digiassess.aspx> (accessed 7 June 2013).
- Jordan, S. (2007) The mathematical misconceptions of adult distance-learning science students. In *Proceedings of the CETL-MSOR Conference, Sept. 2006*, pp87–92. Birmingham: Maths, Stats & OR Network.
- Jordan, S. (2011) Using interactive computer-based assessment to support beginning distance learners of science, *Open Learning* **26** (2), 147–164.
- Jordan, S. (2012a) Short-answer e-assessment questions: five years on. In *Proceedings of CAA 2012 International Conference, Southampton* (ed. D. Whitelock, W. Warburton, G. Wills and L. Gilbert).
- Jordan, S. (2012b) Student engagement with assessment and feedback: some lessons from short-answer free-text e-assessment questions, *Computers & Education* **58** (2), 818–834.
- Jordan, S. (2013) Using e-assessment to learn about learning. In *Proceedings of CAA 2013 International Conference, Southampton*. (ed. D. Whitelock, W. Warburton, G. Wills and L. Gilbert).
- Jordan, S. and Butcher, P. (2010) Using e-assessment to support distance learners of science. In *Physics Community and Cooperation: Selected Contributions from the GIREP-EPEC and PHEC 2009 International Conference* (ed. D. Raine, C. Hurkett and L. Rogers). Leicester: Lula/The Centre for Interdisciplinary Science.
- Jordan, S., Butcher, P. and Ross, S. (2003) *Mathematics assessment at a distance*. Maths CAA Series. Available at <http://ltsn.mathstore.ac.uk/articles/maths-caa-series/july2003/index.shtml> (Accessed 7 June 2013).
- Jordan, S., Jordan, H and Jordan, R. (2012) Same but different, but is it fair? An analysis of the use of variants of interactive computer-marked questions. *International Journal of e-Assessment* **2** (1).

- Jordan, S. and Mitchell, T. (2009) E-assessment for learning? The potential of short-answer free-text questions with tailored feedback. *British Journal of Educational Technology* **40** (2), 371–385.
- Judson, E. and Sawada, D. (2002) Learning from past and present: Electronic response systems in college lecture halls. *Journal of Computers in Mathematics and Science Teaching* **21** (2), 167–181.
- Kay, R.H. and LeSage, A. (2009) Examining the benefits and challenges of using audience response systems: a review of the literature. *Computers & Education* **53** (3), 819–827.
- Kleeman, J. (2013) *Assessment prior art*. Available at <http://assessmentpriorart.org/> (Accessed 7 June 2013).
- Knight, P. (ed.) (1995) *Assessment for learning in higher education*. London: Kogan Page.
- Kornell, N. and Bjork, R.A. (2007) The promise and perils of self-regulated study. *Psychonomic Bulletin & Review* **14**, 219–224.
- Kuechler, W. and Simkin, M. (2003) How well do multiple choice tests evaluate student understanding in computer programming classes. *Journal of Information Systems Education* **14** (4), 389–399.
- Landauer, T.K., Laham, D. and Foltz, P. (2003) Automatic essay assessment. *Assessment in Education* **10**, 295–308.
- Lasry, N., Mazur, E., and Watkins, J. (2008) Peer instruction: from Harvard to the two-year college. *American Journal of Physics* **76** (11), 1066–1069.
- Leacock, C. and Chodorow, M. (2003) C-rater: automated scoring of short-answer questions. *Computers & Humanities* **37** (4), 389–405.
- Lilley, M., Barker, T. and Britton, C. (2004) The development and evaluation of a software prototype for computer-adaptive testing. *Computers & Education* **43** (1), 109–123.
- Littauer, R. (1972) Instructional implications of a low-cost electronic student response system. *Educational Technology: Teacher and Technology Supplement* **12** (10), 69–71.
- Lobb, R. (2013) *Coderunner*. Available at <https://github.com/trampgeek/CodeRunner> (Accessed 7 June 2013).
- Lunt, T. and Curran, J. (2010) 'Are you listening please?' The advantages of electronic audio feedback compared to written feedback. *Assessment & Evaluation in Higher Education* **35** (7), 759–769.
- Luxton-Reilly, A. (2009) A systematic review of tools that support peer assessment. *Computer Science Education* **19** (4), 209–232.
- Luxton-Reilly, A. and Denny, P. (2010) Constructive evaluation: a pedagogy of student-contributed assessment. *Computer Science Education* **20** (2), 145–167.
- McAllister, D. and Guidice, R.M. (2012) This is only a test: A machine-graded improvement to the multiple-choice and true-false examination. *Teaching in Higher Education* **17** (2), 193–207.
- McGarvey, D.J. and Haxton, K.J. (2011) Using audio for feedback on assessments: Tutor and student experiences. *New Directions in the Teaching of Physical Sciences* **7**, 5–9.
- McGuire, G.R., Youngson, M.A., Korabinski, A.A. and McMillan, D. (2002) Partial credit in mathematics exams: A comparison of traditional and CAA exams. In *Proceedings of the 6th International Computer-Assisted Assessment Conference, University of Loughborough*.
- Mackenzie, D. (1999) Recent developments in the Tripartite Interactive Assessment Delivery system (TRIADs). In *Proceedings of the 3rd International Computer-Assisted Assessment Conference, University of Loughborough*.
- Mackenzie, D. (2003) Assessment for e-learning: what are the features of an ideal e-assessment system? In *Proceedings of the 7th International Computer-Assisted Assessment Conference, University of Loughborough*.
- Madden, T. (2007) Supporting student e-portfolios. *New Directions in the Teaching of Physical Sciences* **3**, 1–6.
- Marsh, E.J., Roediger, H.L.III, Bjork, R.A. and Bjork, E.L. (2007) The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review* **14**, 194–199.
- Masters, K. (2011) A brief guide to understanding MOOCs. *The Internet Journal of Medical Education* **1** (2).
- Mathews, J. (2006) Just whose idea was all this testing? *The Washington Post*, 14 November 2006.
- Mazur, E. (1991) *Peer instruction: A user's manual*. New Jersey: Prentice-Hall.
- Millar, J. (2005) *Engaging students with assessment feedback: what works? An FDTL5 Project literature review*. Oxford: Oxford Brookes University.
- Miller, T. (2008) Formative computer-based assessment in higher education: the effectiveness of feedback in supporting student learning. *Assessment & Evaluation in Higher Education* **34** (2), 181–192.
- Mitchell, T., Aldridge, N., Williamson, W. and Broomhead, P. (2003) Computer based testing of medical knowledge. In *Proceedings of the 7th*

International Computer-Assisted Assessment Conference, University of Loughborough.

Mitchell, T., Russell, T., Broomhead, P. and Aldridge, N. (2002) Towards robust computerised marking of free-text responses. In *Proceedings of the 6th International Computer-Assisted Assessment Conference, University of Loughborough.*

Molyneaux, T., Brumley, J., Li, J., Gravina, R. and Ward, L. (2009) A comparison of the use of Wikis and ePortfolios to facilitate group project work. In *Proceedings of the 20th Annual Conference for the Australasian Association for Engineering Education: Engineering the Curriculum, Adelaide, Australia* (ed. C. Kestell, S. Grainger and J. Cheung), pp388–393.

Moodle (2013) Available at <https://moodle.org/about/> (accessed 7 June 2013).

Naismith, L. and Sangwin, C.J. (2004) Computer algebra based assessment of mathematics online. In *Proceedings of the 8th International Computer-Assisted Assessment Conference, University of Loughborough.*

Nicol, D. (2007) E-assessment by design: Using multiple choice tests to good effect. *Journal of Further & Higher Education* **31** (1), 53–64.

Nicol, D. (2008) *Technology-supported assessment: A review of research.* Unpublished manuscript available at <http://www.reap.ac.uk/resources.html> (Accessed 7th June 2013).

Nicol, D.J. and Macfarlane-Dick, D. (2006) Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education* **31** (2), 199–218.

Nix, I. and Wyllie, A. (2011) Exploring design features to enhance computer-based assessment: Learners' views on using a confidence-indicator tool and computer-based feedback. *British Journal of Educational Technology* **42** (1), 101–112.

O'Malley, P.J. (2011) Combining screencasting and a tablet PC to deliver personalised student feedback. *New Directions in the Teaching of Physical Sciences* **7**, 27–30.

Orrell, J. (2008) Assessment beyond belief: the cognitive process of grading. In *Balancing dilemmas in assessment and learning in contemporary education* (ed. A. Havnes and L. McDowell), pp251–263. New York & Oxford: Routledge.

Parkin, H.J., Hepplestone, S., Holden, G., Irwin, B. and Thorpe, L. (2012) A role for technology in enhancing students' engagement with feedback. *Assessment & Evaluation in Higher Education* **37** (8), 963–973.

Pearson (2013) MasteringPhysics®. Available at <http://www.masteringphysics.com/site/index.html> (accessed 7 June 2013).

Pebblepad (2013) Available at <http://www.pebblepad.co.uk/> (accessed 7 June 2013).

PeerWise (2013) Available at <http://peerwise.cs.auckland.ac.nz/> (accessed 7 June 2013).

Perelman, L. (2008) Information illiteracy and mass market writing assessments. *College Composition and Communication* **60** (1), 128–141.

Purchase, H., Hamer, J., Denny, P. and Luxton-Reilly, A. (2010) The quality of a PeerWise MCQ repository. In *Proceedings of the Twelfth Australasian Conference on Computing Education*. pp137–146.

Pyper, A. and Lilley, M. (2010) A comparison between the flexilevel and conventional approaches to objective testing. In *Proceedings of CAA 2010 International Conference, Southampton* (ed. D. Whitelock, W. Warburton, G. Wills and L. Gilbert).

Rebello, N.S. and Zollman, D.A. (2004) The effect of distracters on student performance on the force concept inventory. *American Journal of Physics* **72**, 116–125.

Redecker, C., Punie, Y. and Ferrari, A. (2012) eAssessment for 21st Century Learning and Skills. In *21st Century Learning for 21st Century Skills, Proceedings of the 7th European Conference of Technology Enhanced Learning, Saarbrücken, Germany* (eds. A. Ravenscroft, S. Lindstaedt, C.D. Kloos and D. Hernandez-Leo).

Ridgway, J., McCusker, S. and Pead, D. (2004) *Literature review of e-assessment.* Futurelab report 10. Bristol: Futurelab.

Ripley, M (2007) *E-assessment: An update on research, policy and practice.* Bristol: Futurelab.

Roediger, H.L. III, and Karpicke, J.D. (2006) The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science* **1** (3), 181–21.

Roediger, H.L. III, and Marsh, E.J. (2005) The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, & Cognition* **31**, 1155–1159.

Rosewell, J.P. (2011) Opening up multiple-choice: assessing with confidence. In *Proceedings of CAA 2011 International Conference, Southampton.* (ed. D. Whitelock, W. Warburton, G. Wills and L. Gilbert).

Ross, S., Jordan, S. and Butcher, P. (2006) Online instantaneous and targeted feedback for remote learners. In *Innovative Assessment in Higher Education* (ed. C. Bryan and K. Clegg). London: Routledge. pp123–131.

Sandeen, C. (2013) Assessment's place in the new MOOC world, *Research & Practice in Assessment* **8**, 5–12.

- Sangwin, C.J. (2013) *Computer aided assessment of mathematics*. Oxford: Oxford University Press.
- Sangwin, C.J. and Grove, M.J. (2006) STACK: Addressing the needs of the 'neglected learners'. In *Proceedings of the First WebALT Conference and Exhibition, Technical University of Eindhoven, Netherlands*, pp81–95.
- SCHOLAR (2013) Available at scholar.hw.ac.uk (accessed 13 August 2013).
- Scouller, K. (1998) The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay, *Higher Education* **35**, 453–472.
- Shaul, M. (2007) Assessing online discussion forum participation. *International Journal of Information and Communication Technology Education* **3** (3), 39–46.
- Siddiqi, R. and Harrison, C.J. (2008) On the automated assessment of short free-text responses. In *Proceedings of the 34th International Association for Educational Assessment (IAEA) Annual Conference, Cambridge*.
- Sim, J.W.S. and Hew, K.F. (2010) The use of weblogs in higher education settings: A review of empirical research. *Educational Research Review* **5** (2), 151–163.
- Simkin, M.G. and Kuechler, W.L. (2005) Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education* **3** (1), 73–97.
- Simpson, V. and Oliver, M. (2007) Electronic voting systems for lectures then and now: A comparison of research and practice. *Australasian Journal of Educational Technology* **23** (2), 187–208.
- Stödborg, U. (2012) A research review of e-assessment. *Assessment & Evaluation in Higher Education* **37** (5), 591–604.
- Strickland, N. (2002) Alice Interactive Mathematics, *MSOR Connections* **2** (1), 27–30.
- Sukkarieh, J.Z., Pulman, S.G. and Raikes, N. (2003) Auto-marking: using computational linguistics to score short, free-text responses. In *Proceedings of the 29th International Association for Educational Assessment (IAEA) Annual Conference, Manchester*.
- Sukkarieh, J.Z., Pulman, S.G. and Raikes, N. (2004) Auto-marking 2: using computational linguistics to score short, free-text responses. In *Proceedings of the 30th International Association for Educational Assessment (IAEA) Annual Conference, Philadelphia*.
- Thomas, R. and Milligan, C. (2003) Online assessment of practical experiments. In *Proceedings of the 7th International Computer-Assisted Assessment Conference, University of Loughborough*.
- TRIADS. Available at <http://www.triadsinteractive.com/> (accessed 7th June 2013).
- Valenti, S., Neri, S. and Cucchiarelli, A. (2003) An overview of current research on automated essay grading. *Journal of Information Technology Education* **2**, 319–330.
- Van Labeke, N., Whitelock, D., Field, D., Pulman, S. and Richardson, J.T.E. (2013) OpenEssayist: extractive summarisation and formative assessment of free-text essays. In *Proceedings of the 1st International Workshop on Discourse-Centric Learning Analytics, Leuven, Belgium*.
- Veloski, J.J., Rabinowitz, H.K., Robeson, H.R. and Toung, P.R. (1999) Patients don't present with five choices: an alternative to multiple-choice tests in assessing physicians' competence. *Academic medicine* **74** (5), 539–546.
- Ventouras, E., Triantis, D., Tsiakas, P and Stergiopoulos, C. (2010) Comparison of examination methods based on multiple-choice questions and constructed-response questions using personal computers. *Computers & Education* **54**, 455–461.
- Vojak, C., Kline, S., Cope, B., McCarthey, S. and Kalantzis, M. (2011) New spaces and old places: An analysis of writing assessment software. *Computers & Composition* **28**, 97–111.
- Walet, N. and Birch, M. (2012) Using online assessment to provide instant feedback. In *Proceedings of the HEA-STEM Annual Conference, London*.
- Whitelock, D. and Brasher, A. (2006) Roadmap for e-assessment: Report for JISC. Available at http://www.jisc.ac.uk/elp_assessment.html#downloads (accessed 7th June 2013).
- Wieman, C. (2010) Why not try a scientific approach to science education? *Change: The Magazine of Higher Learning* **39** (5), 9–15.
- Wyllie, A. and Waights, V. (2010) CDM. Available at <https://students.open.ac.uk/openmark/cdm.projectworld/> (accessed 7 June 2013).