Assessment for learning: pushing the boundaries of computer-based assessment

Practitioner Research
in Higher Education

Vol 3 (1): page 11-19

Sally Jordan
The Open University
s.e.jordan@open.ac.uk

Abstract
Feedback on assessment tasks has an important part to play in underpinning student learning. Online assessment enables instantaneous feedback to be given so that the student can act on it immediately. However, concern has been expressed that e-assessment tasks (especially multiple-choice questions) can encourage surface-learning. Several projects at the UK Open University are investigating the use of rich interactive e-assessment. One of these projects is using a linguistically based authoring tool to enable sophisticated answer matching for free-text responses of up to a sentence in length. Immediate tailored feedback is provided on incorrect and incomplete responses, and students are able to use this feedback in reattempting the question. Students have been observed attempting the questions and were seen to answer them in different ways, with most students using short phrases but some using full and carefully constructed sentences and some using note form. There was evidence that some students assumed the system to be looking only for keywords. A human-computer marking comparison has demonstrated the computer-based answer matching to be of similar or greater accuracy than that of six course tutors.

Keywords
Assessment for learning; computer-based assessment; feedback; short-answer questions; natural language processing.

Introduction
Much has been written about the impacts, positive and negative, intentional or otherwise, that assessment practice has on students' learning. Assessment has been identified as the 'single biggest influence on how students approach their learning' (Rust et al., 2005:231). Distinctions are frequently made between so-called formative and summative assessment, but given the inevitable driver of grading, Barnett (2007:37) rightly notes that:

> *Summative assessment is itself formative… at issue is whether that formative potential of summative assessment is lethal or emancipatory.*

Assessment can lead students to concentrate on certain topics (i.e. it can define what students study); it can also alter students' learning approaches (and so define how the studying is done) (Scouller and Prosser, 1994). This is not a new effect. Snyder (1971) identified the dissonance between the formal curriculum and the 'hidden curriculum', driven by the hurdles (including examinations and other assessment tasks) that students perceive they are required to jump. Recently, a group of leading academics and assessment experts (Weston Manor Group, 2008) have called for a change in assessment priorities, in an attempt to place a greater focus on assessment for learning rather than assessment of learning, and in particular to free students from the obsession with marks which is seen as encouraging them to adopt a strategic approach to their studies.

Reviews of the literature (e.g. Black and Wiliam, 1998; Gibbs and Simpson, 2004) have identified conditions under which assessment appears to support and encourage learning. These have been developed into a number of frameworks, to be used by practitioners in developing and auditing assessment practice (Gibbs and Simpson, 2004; Nicol and Macfarlane-Dick, 2006). Not surprisingly these frameworks share common themes, centred around assessment's power to engage and motivate students and the role of feedback in helping students to improve. However, the provision of feedback does not in itself lead to learning. Sadler (1989:119) reports the

> *… common but puzzling observation that even when teachers provide students with valid and reliable judgments about the quality of their work, improvement does not necessarily follow.*

Sadler argues that in order for feedback to be effective, action must be taken to close the gap between the student's current level of understanding and the level expected by the teacher. In taking this view he is aligning himself with Ramaprasad (1983), going beyond a definition of feedback as purely the transmission of information from teacher to learner, to one in which the information must be used to alter the gap. This is in line with the scientific definition of feedback as a cyclical process, in which a change in one parameter leads to a change in the initial conditions.

It follows that, in order for assessment to be effective, feedback must not only be provided, but also understood by the student and acted on in a timely fashion. These points are incorporated into five of Gibbs and Simpson's (2004) eleven conditions under which assessment supports learning:

Condition 4: Sufficient feedback is provided, both often enough and in enough detail.
Condition 6: The feedback is timely in that it is received by students while it still matters to them and in time for them to pay attention to further learning or receive further assistance.
Condition 8: Feedback is appropriate, in relation to students' understanding of what they are supposed to be doing.
Condition 9: Feedback is received and attended to.
Condition 11: Feedback is acted upon by the student.

A role for e-assessment?

It can be difficult and expensive for teachers to provide their students with sufficient feedback (Condition 4), especially if students are studying part-time or are in a distance-learning environment, so opportunities for informal discussion are limited. Pressure of work can lead teachers to return feedback when it is too late to be useful (Condition 6) and it is then difficult for students to understand and act upon it (Conditions 8 and 10), even assuming that they bothered to collect the work and to do more than glance at the mark awarded (Condition 9).

One possible solution to these dilemmas is to use e-assessment. Feedback can be tailored to students' misconceptions and delivered instantaneously and, provided the assessment system is carefully chosen and set up, students can be given an opportunity to learn from the feedback while it is still fresh in their minds, by immediately attempting a similar question or the same question for a second time, thus closing the feedback loop. Part-time and distance learners are no longer disadvantaged and 'little and often' assessments can be incorporated at regular intervals throughout the module, bringing the additional benefits of assisting students to pace their study and to engage actively with the learning process, thus encouraging retention. For high-population modules and programmes, e-assessment can also deliver savings of cost and effort. Finally, e-assessment is the natural partner to the growth industry of e-learning.

There is some optimism and excitement about the possibilities offered (Whitelock and Brasher, 2006). However opinions of e-assessment are mixed and evidence for its effectiveness is inconclusive; indeed e-assessment is sometimes perceived as having a negative effect on learning (Gibbs, 2006). Murphy (2008) reports that high stakes multiple-choice tests of writing can lead to actual writing beginning to disappear from the curriculum; she also reports that 'the curriculum begins to take the form of the test' (2008:36). There are more widely voiced concerns that e-assessment tasks (predominantly but not exclusively multiple-choice) can encourage memorisation and factual recall and lead to surface-learning, far removed from the tasks that will be required of the learners in the real world (Mitchell et al., 2003; Scouller and Prosser, 1994). Also, although multiple-choice questions are in some senses very reliable, they may not always be assessing what the teacher believes that they are, partly because multiple-choice questions require 'the recognition of the answer rather than the construction of a response' (Nicol, 2007:54).

Ashton and her colleagues (2006) point out that the debate about the effectiveness of multiple-choice questions can divert attention away from many of the benefits that online assessment can offer to learning. Perhaps the question we should be asking is not 'should we be using e-assessment?' but rather 'what can we do to make e-assessment more effective?'

E-assessment at the Open University

The work described here is part of an initiative at the UK Open University that is studying the effectiveness of rich interactive computer-marked assessment and feedback in promoting student learning. The students in question are all distance learners and they are mostly part-time. Thus the challenges of providing timely and useful feedback, described above, are at their most severe. The initiative is organised within a practitioner-led action research framework, with each practitioner working on a separate project to investigate a novel approach to or application of e-assessment.

The interactive computer-marked assessment (iCMA) initiative is part of the wider university's e-assessment activity. Most of the iCMA projects make use of the OpenMark assessment system (Marshall, 2008), also used in a number of mainstream settings across the university. OpenMark operates within the Moodle virtual learning environment and information about a student's progress through an e-assessment activity (also known as an iCMA) can be recorded and passed to the student's course tutor, enabling appropriate support to be offered. OpenMark incorporates a number of question types, allowing for the free-text entry of numbers, simple algebraic expressions and single words as well as drag-and-drop, hotspot, multiple choice and multiple response questions.

A feature of the OpenMark system is that students are allowed multiple attempts at each question before proceeding, with the amount of feedback provided increasing at each attempt. If the questions are used summatively, the mark awarded decreases at each attempt, but the presence of multiple attempts with increasing feedback remains a feature. Thus, even in use that is technically summative, the focus is on assessment for learning. At the first attempt, an incorrect response will usually result in the simple feedback 'Your answer is incorrect', which gives the student the opportunity to correct their answer with the minimum of assistance.

If the student's response is still incorrect at the second attempt, they will receive a more detailed hint, with a reference to the course material. At the third (final) attempt, the student will receive a complete answer, again with a reference to the course material. Whenever possible, the feedback is targeted to the misunderstanding that has led to the error. The provision of multiple attempts with increasing feedback is designed to give the student an opportunity to correct his or her work immediately (i.e. to act on the feedback provided – Gibbs and Simpson Condition 11) and the tailored feedback is designed to simulate a 'tutor at the student's elbow', offering feedback that is as appropriate and helpful as possible (Ross et al., 2006).

Each question exists in several variants, chosen to be of similar difficulty. In purely formative use, this provides extra practice; in summative use it reduces opportunities for plagiarism. Students can access the iCMA as frequently and for as long as they would like to (in summative use this is typically within a window of several weeks; in formative use it is for the duration of the module), from any computer supporting the Firefox or Internet Explorer browsers, and the system records their progress. Questions can be attempted in any order, although students are encouraged to do them in the order offered.

Short answer free-text questions with feedback

In an attempt to extend the application of e-assessment of this type, a pilot study is using an authoring tool provided by Intelligent Assessment Technologies Ltd. (IAT) (Mitchell et al., 2002) to write questions requiring free-text answers of up to around 20 words in length. The authoring tool uses the natural language processing (NLP) technique of information extraction, and incorporates a number of processing modules aimed at providing accurate marking without undue penalty for errors in spelling and grammar. The question authors are not NLP or programming experts, but use an interface to the authoring tool which enables mark schemes to be represented as a series of templates.

The linguistically-based answer matching means that it is possible to accurately mark many different and sometimes quite complex student responses (see Figure 1). It is possible to distinguish an answer such as 'Kinetic energy is converted into gravitational energy' from 'Gravitational energy is converted into kinetic energy' and the negated form of a correct response can be marked as incorrect (so 'The forces are balanced' is marked as correct while 'The forces are not balanced' is marked as incorrect).

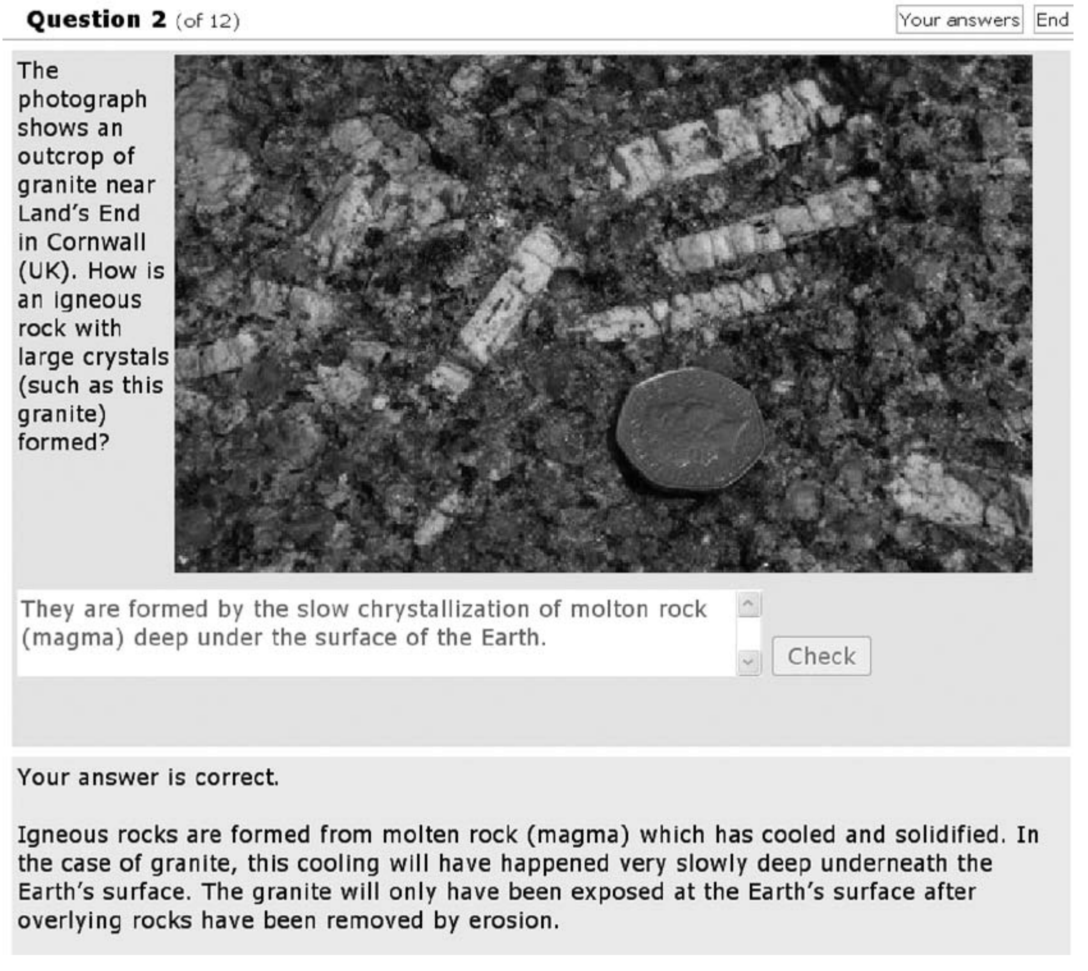

Figure 1 A free-text question, illustrating the correct marking of a complex answer

A novel feature of the current project has been the use of student responses to early developmental versions of the questions – themselves delivered online – to improve the answer matching. Previous users of similar software (e.g. Mitchell et al., 2003; Sukkarieh et al., 2003) have used student responses to paper-based questions to develop the computer-based answer matching, but this approach assumes that there are no characteristic differences between student responses to the same question delivered by different media, or between responses that students assume will be marked by a computer as opposed to a human marker.

A second novel feature of the work is the emphasis placed instantaneous targeted feedback. The questions are offered to students via OpenMark, so students are allowed several attempts as described above. However the feedback for incomplete or incorrect answers (as shown in Figure 2) is generated from within the IAT authoring tool. Targeted feedback has been added for misconceptions and omissions observed in the analysis of student responses.

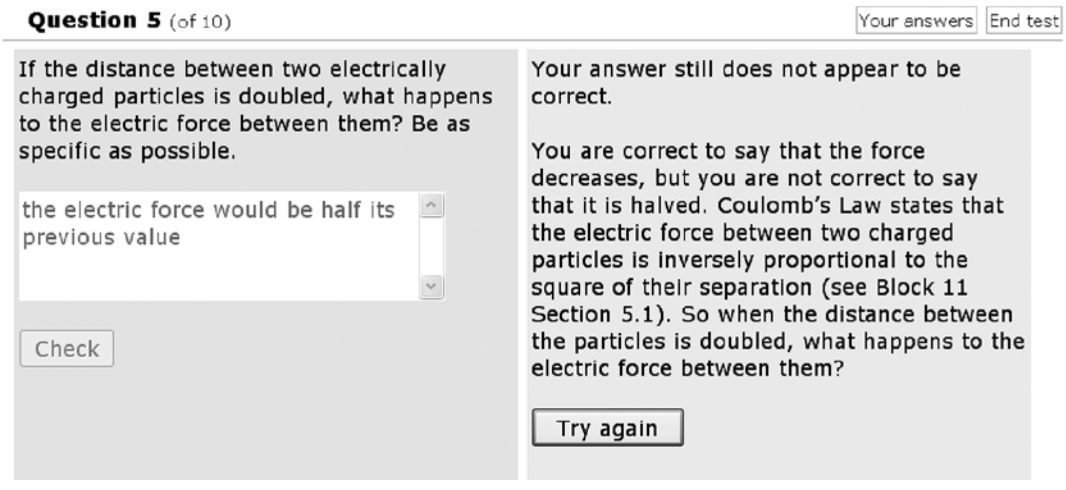

Figure 2 A free-text question, showing targeted feedback on an incorrect answer

Seventy-five short-answer questions, assessing the learning outcomes of an introductory interdisciplinary science course, have been authored and refined in the light of students' responses. Evaluation has focused on student reaction to questions of this type, their use of the feedback provided and on the accuracy of marking relative to that of human markers.

Modified versions of the some of the questions developed have been incorporated, along with conventional OpenMark questions, into regular iCMAs which form part of an integrated assessment policy (also including tutor-marked assessment) for a new module. These iCMAs are summative, but low stakes; their role is to encourage students to keep up to date in their studies as well as providing instantaneous tailored feedback and an opportunity for students to act on that feedback immediately.

Evaluation 1: Student reaction and use of feedback

In order to evaluate the usability of the questions in a controlled setting, six student volunteers were observed in the OU's Institute for Educational Technology Usability Laboratory. The students were asked to attempt one of two iCMAs, each including a number of free-text questions alongside a number of conventional OpenMark questions. In the following discussion all student names have been altered.

In line with accepted practice for usability laboratory observation (see for example Stone et al., 2005), participants interacted with the iCMA without assistance. The participants' interaction with the questions was observed live and recorded for subsequent analysis (Figure 3). A verbal think-aloud protocol was used, whereby the participants were asked to talk about what they were doing and thinking, and after the evaluation session itself, each participant was asked to comment retrospectively on the reasons for their actions and on the reaction to the different question types. Analysis of the recordings is in progress; early results are reported here, alongside findings from an analysis of student responses to all the questions and informal feedback, gathered from feedback questions at the end of each iCMA and from an online forum.

Figure 3 A screen-shot from a recorded usability laboratory session

Students were not initially told anything about the technology behind the questions; some asked for more information, others were clearly experimenting for themselves (e.g. Philip 'What's the minimum you can put in? If I put absorbed? [he then typed a single word and it was marked as correct]). Most were very impressed by the answer-matching. Some students reported specific questions in which they considered their response to have been inaccurately marked, usually where they had been marked as incorrect. In many of these cases their responses were indeed incorrect and sometimes targeted feedback had highlighted the specific nature of the student's misunderstanding. Some students said they would prefer multiple-choice questions, because:

> *… in multiple choice, obviously you know that the answer is there somewhere, it's just a matter of finding it, so there is an element of I'm not going to be completely out.*

Five of the six students observed in the laboratory entered their answers as phrases rather than complete sentences, but one student, Colin, entered very complete answers and checked them carefully, making adjustments to the word order and punctuation before submitting his answer for marking. So while most students answer the first question with a phrase such as 'coloured lines', Colin's final answer, after re-reading his answer several times and altering word order, was 'The spectrum will be characterised by a vertical line showing where in the spectrum the particular colour is absorbed by the vapour.' The length of Colin's answers were initially assumed to be evidence that he was putting in as many keywords as possible in an attempt to match the required ones, but the careful phrasing of his answers makes this explanation seem unlikely; Colin started off by commenting that he was 'going to answer the questions in the same way as for a tutor-marked assignment' and it appears that he was doing just that.

Students were not initially given any indication of the form of answer expected; latterly they were advised to enter their answers as a simple sentence, but most continued to enter phrases rather than complete sentences. It is not clear whether students were doing this because they were assuming that the computer's marking was simply keyword-based, or because the question was written immediately above the answer so they felt there was no need to repeat words from the question in the first part of their answer. A small number of responses show evidence of students trying to 'help' the computer by entering very terse answers, for example 'fragmental. permeable. porous'. In addition, when their first one or two attempts had been marked as incorrect, some students simply added additional words, which resulted in an overall answer that did not make sense, presumably in the hope that the extra words would match the required answer. So 'Banding of different materials, small grain size and no crystals' became 'Banding of different materials, small grain size and no crystals sand cement'.

Students also appeared to use the feedback in different ways. Some of the students were observed to reading the feedback carefully and act on it. Julia scrolled across the screen so as to be able to read all of the feedback provided, read out parts of it, nodded and said 'OK' to indicate that she had understood it. When told that an answer was incorrect and given targeted feedback she read aloud:

> *You are on the right lines but you need to specify how much further apart the energy levels are…*

then said 'fair enough' and went back to two previous questions that she (rightly) assumed would help her to work out the answer to this question – and she got it right at the next attempt. Similarly, Malcolm used targeted feedback to find the right section in the book and so to amend his answer. However, evidence that students do not always read written feedback carefully came from instances where an incorrect answer was marked as correct. For example, Colin's careful answer to the first question (given above) was actually incorrect, but unfortunately the computer marking was too loose and it was marked as correct. Colin appeared to read the question author's answer (which he received immediately after he had given his response) but he did not appear to notice that this was at variance with the answer he had entered. It seems likely that others, like Charlotte, reasoned

> *… if I got it right and thought I had the process right I didn't always read the answers.*

So being told that an incorrect answer is correct might be acting to reinforce previous misunderstanding.

Evaluation 2: Human-computer marking comparison

Between 92 and 246 student responses to each of seven free-text questions were marked independently by the computer system, by six course tutors and by the question author.

To ensure that the human-computer marking comparison did not assume that either the computer or the human markers were 'right', the IAT and each course tutor's marking of each response were compared against:

- the median of all the course tutors' marks for that response
- the 'blind' marking of the response by the author of the questions.

Responses in which there was any divergence between the markers and/or the computer system were inspected in more detail, to investigate the reasons for the disagreement.

Chi-squared tests showed that, for three of the questions, the marking of all the markers (including the computer system) was indistinguishable. For the other four questions, the markers were marking in a way that was significantly different. However, in all cases, the mean mark allocated by the computer system was within the range of means allocated by the human markers. In some cases the differences between human markers were large – for Question 13

> *You are handed a rock specimen from a cliff that appears to show some kind of layering. The specimen does not contain any fossils. How could you be sure, from its appearance, that this rock specimen was a sedimentary rock?*

the mean mark awarded by the most lenient tutor was 2.5 times the mean mark awarded by the most severe.

Analysis of variance indicated that overall marking of the markers fell into two distinct groups, but the computer marking was consistent with the majority of the human markers. For individual questions, the percentage of responses where there was any variation in marking varied between 4.8% (for Question 1; in which the word 'direction' was an adequate response) and 64.4% (for Question 13) but in every case more variation was caused by discrepancy between the course tutors than between the median of the course tutors or the question author and the computer system.

For six of the questions the marking of the computer system was in agreement with that of the question author for more than 95% of the responses (rising as high as 99.5% for Question 1). For Question 13, the least well developed of the questions at the time the comparison took place, there was agreement with the question author for 87.4% of the responses. Improvements to the answer matching since the human-computer marking comparison took place would result in 97.0% agreement now.

Mitchell et al. (2002) identified the following reasons for inaccurate computer marking:
• omission of a mark scheme template
• failure to correctly identify miss-spelled or incorrectly used words
• failure to properly analyse the sentence structure
• failure to identify an incorrect qualification (where a correct response is nullified by an incorrect one).

In the current analysis there were examples of each of these, but all were relatively rare and the first three were not considered to be significant issues. However the final reason for inaccurate computer marking, where for example the computer marked the response 'direction and acceleration' as correct because of its mention of 'direction', whereas the question author and the course tutors all felt that the mention of 'acceleration' made it clear that the student did not demonstrate the relevant knowledge and understanding learning outcome, represents a serious threat to the accuracy of any computer marking of free-text answers. While any individual incorrect response of this nature can be dealt with (in the IAT authoring tool by the addition of a 'do not accept' mark-scheme) it is not realistic to make provision for all flawed answers of this type.

Reflection
Inaccuracy of human marking has been identified as a concern by Orrell (2008) and the Office of the Qualifications and Examinations Regulator (reported by Frean, 2008), and this study has demonstrated that computers can mark short-answer free-text questions as accurately as human markers. When the concern is with assessment for learning rather than the assessment of learning, perhaps the accuracy of marking should not matter too much, but if marks are used to encourage students to engage with the assessment task, they will inevitably be concerned about the accuracy of the marking and they are likely to have less confidence in computers than human markers. Rightly or wrongly, students are also likely to have less confidence in free-text marking than they have in the marking of multiple-choice questions. A solution to the perennial problem of marks 'getting in the way' of teaching might be to make the iCMAs compulsory but not scoring; the problem then is that students are unlikely to engage with the iCMAs or the feedback provided in as serious a way.

Accuracy of marking remains important because of the importance of giving correct feedback to students, in particular not telling them that an incorrect answer is correct. The finding that different students make very different uses of feedback is in line with McDowell's (2008) findings about students' varied use of feedback on more conventional assessment tasks.

Whitelock and Brasher (2006) identified the principal barriers to the development of institution-wide e-assessment as being staff time and training. Learning how to use a linguistically-based authoring tool is undoubtedly time-consuming, and the writing of good e-assessment questions and embedding them within an appropriate assessment strategy that truly supports student learning are skills that should not be underestimated.

Nevertheless, the author believes that carefully designed online interactive assessment can improve the learning environment for students across a range of disciplines and institutions. E-assessment is not a panacea, but the work reported here demonstrates the potential of rich e-assessment tasks to support student learning.

Acknowledgements
The financial support of the Centre for Open Learning of Mathematics, Science, Computing and Technology (COLMSCT) is gratefully acknowledged, as is the assistance of many members of COLMSCT staff, especially Barbara Brockbank and Philip Butcher and Tom Mitchell of Intelligent Assessment Technologies Ltd.

References
All websites accessed 17.12.08
Ashton, H.S., Beevers, C.E., Milligan, C.D. et al. (2006) Moving beyond objective testing in online assessment. In S.C. Howell and M. Hricko (eds) *Online Assessment and Measurement: Case Studies From Higher Education, K-12 and Corporate.* Hershey, PA: Information Science Publishing, 116–127.
Barnett, R. (2007) Assessment in higher education:an impossible mission?. In D. Boud and N. Falchikov (eds) *Rethinking Assessment in Higher Education.* London: Routledge, 29–40.
Black, P. and Wiliam, D. (1998) Assessment and classroom learning, *Assessment in Education,* 5(1): 7–74.
Frean, A. (2008) GCSE and A level exam results inaccurate, Ofqual warns pupils. www.timesonline.co.uk/tol/life_and_style/education/article3941601.ece
Gibbs, G. (2006) Why assessment is changing. In C. Bryan and K.V.Clegg (eds) *Innovative Assessment in Higher Education.* London: Routledge, 11–22.

Gibbs, G. and Simpson, C. (2004) Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education* 1: 3–31.
Marshall, S. (2008) *OpenMark* question system. https://openmark.dev.java.net/
McDowell, L. (2008) Students' experiences of feedback on academic assignments in higher education: implications for practice. In A. Havnes and L. McDowell (eds) *Balancing Dilemmas in Assessment and Learning in Contemporary Education.* London: Routledge, 237–249.
Mitchell, T., Aldridge, N., Williamson, W. and Broomhead, P. (2003) Computer-based testing of medical knowledge. 7th International CAA Conference, Loughborough, UK. www.caaconference.com/pastconferences/2003/procedings/mitchell.pdf
Mitchell, T., Russell, T., Broomhead, P. and Aldridge, N. (2002) Towards robust computerised marking of free-text responses. 6th International CAA Conference, Loughborough, UK. www.caaconference.com/pastconferences/2002/proceedings/Mitchell_t1.pdf
Murphy, S. (2008) Some consequences of writing assessment. In A. Havnes and L. McDowell (eds) *Balancing Dilemmas in Assessment and Learning in Contemporary Education.* London: Routledge, 33–49.
Nicol, D.J. (2007) E-assessment by design: using multiple choice tests to good effect. *Journal of Further and Higher Education.* 31(1): 53–64.
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.
Orrell, J. (2008) Assessment beyond belief: the cognitive process of grading. In A. Havnes and L. McDowell (eds) *Balancing Dilemmas in Assessment and Learning in Contemporary Education.* London: Routledge, 251–263.
Ramaprasad, A. (1983) On the definition of feedback. *Behavioral Science* 28: 4–13.
Ross, S.M., Jordan, S.E and Butcher, P.G. (2006) Online instantaneous and targeted feedback for remote learners in C. Bryan and K.V. Clegg (eds) *Innovative Assessment in Higher Education.* London: Routledge, 123–131.
Rust, C., O'Donovan, B. and Price, M. (2005) A social constructivist assessment process model: how the research literature shows us this could be best practice. *Assessment & Evaluation in Higher Education* 30(3): 231–240.
Sadler, D.R. (1989) Formative assessment and the design of instructional systems. *Instructional Science* 18: 119–144.
Scouller, K.M. and Prosser, M. (1994) Students' experiences in studying for multiple choice question examinations. *Studies in Higher Education* 19(3): 267–279.
Snyder, B.R. (1971) *The Hidden Curriculum.* New York: Alfred A. Knopf.
Stone, D., Jarratt, C., Woodroffe, M. and Michocha, S. (2005) User interface design and evaluation. San Francisco: Morgan Kaufmann.
Sukkarieh, J.Z., Pulman, S.G. and Raikes, N. (2003) Auto-marking: using computational linguistics to score short, free-text responses. 29th International Association for Educational Assessment (IAEA) Annual Conference, Manchester.
Weston Manor Group (2008) *Assessment standards: a manifesto for change.* www.business.brookes.ac.uk/learningandteaching/aske/Manifesto%20final.pdf
Whitelock, D. and Brasher, A. (2006) *Roadmap for e-assessment.* Joint Information Systems Committee Report. www.jisc.ac.uk/elp_assessment.html