Open Research Online

The Open University's repository of research publications
and other research outputs

Online interactive assessment with short free-text questions and tailored feedback

Journal Item

For guidance on citations see FAQs.

Version: Accepted Manuscript

oro.open.ac.uk

Online interactive assessment with short free-text questions and tailored feedback

Sally Jordan
Centre for the Open Learning of Mathematics, Science, Computing and Technology (COLMSCT)
The Open University
Milton Keynes
MK7 6AA
s.e.jordan@open.ac.uk

Abstract
A linguistically based authoring tool has been used to write e-assessment questions requiring short free-text answers of up to about 20 words in length (typically a single sentence). The answer matching is sophisticated and students are provided with instantaneous targeted feedback on incorrect and incomplete responses. They are able to use this feedback in reattempting the question. Seventy-five questions of this type have been offered to students on an entry-level interdisciplinary science module and they have been well received. Students have been observed attempting the questions and have been seen to respond in differing ways to both the questions themselves and the feedback provided. The answer matching has been demonstrated to be of similar or greater accuracy than specialist human markers.

The software described is all either open source or commercially available, but the purpose of this paper is not to advertise these products but rather to encourage reflection on e-assessment's potential to support student learning.

Introduction
It is widely recognised that assessment has a profound effect on student learning[1], although the effect is not always a positive one[2,3]. There has been a recent explosion of interest in 'assessment for learning'[4], in which the focus is on the formative, teaching functions of assessment and which contrasts with 'assessment of learning', where the primary interest is in the measuring, summative function.

A number of literature reviews[5,6] have identified conditions under which assessment supports student learning[6,7]. Two common themes are assessment's ability to motivate and engage students, and the role of feedback. However, if feedback is to be effective, it must be more than a transmission of information from teacher to learner. The student must understand the feedback sufficiently well to be able to learn from it i.e. to 'close the gap' between their current level of understanding and the level expected by the teacher[8,9]. Thus the use of the word 'feedback' in the context of assessment becomes aligned with the scientific use of the word, as cyclical process in which a change in one variable leads to a change in the initial conditions.

If a student is to be able to act on the feedback received, it follows that, as well as being sufficiently detailed and framed in a way that the student can understood, it must reach them quickly. However it can be difficult for hard-pressed university lecturers to deliver useful feedback in a timely fashion. One possible solution is to use e-assessment; this offers particular benefits when class sizes are large and so development costs are more than offset by savings of academic time. However opinions of e-assessment are mixed: some people are excited about its potential[10]; others are concerned that e-assessment tasks (primarily but not exclusively multiple-choice questions) can encourage a surface approach to learning[3,11]. This paper describes a project, building on work described in Swithenby's review of screen-based assessment[12], which seeks to develop and evaluate more meaningful e-assessment questions.

The work described is taking place at the Open University, as one of a number of 'e-assessment for learning' projects funded by the Centre for the Open Learning of Mathematics, Science, Computing and Technology (COLMSCT)[13]. The software described is all either open source or commercially available, but the purpose of this paper is not to advertise these products but rather to encourage reflection on e-assessment's potential to support student learning.

Background to the current work
The Open University is the global maintainer of the Moodle[14] virtual learning environment quiz engine. Work has been done to improve Moodle question types and reporting, and the OpenMark e-assessment system has been incorporated into Moodle[15]. OpenMark[16] offers 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, hot spot, multiple-choice and multiple-response questions. A distinctive feature of OpenMark is the

provision of multiple attempts at each question, 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. Wherever possible the feedback is tailored to the misunderstanding that has led to the error (see Figure 1). The provision of multiple attempts with increasing feedback is designed to give the student an opportunity to act on the feedback to correct his or her work immediately and the tailored feedback is designed to simulate a 'tutor at the student's elbow'[17].

What is the second derivative of
$z = 3t^3 + 4t^2 - 2t + 3$ with respect to t?

Complete the equation

$\dfrac{d^2z}{dt^2} =$ 9t²-2+8t

☐ Superscript (↑)

Check

Your answer is still incorrect.

You have given the first derivative of z with respect to t. Note that the question asked for the second derivative. See Section 10.2.5.

Try again

Figure 1 Feedback provided on an incorrect response to an OpenMark question.

Short free-text questions with tailored feedback
The current project has extended the range of e-assessment questions offered to students via OpenMark to include those requiring free-text answers of up to around a sentence in length (around 20 words). The answer matching is written with an authoring tool provided by Intelligent Assessment Technologies Ltd. (IAT), which uses the natural language processing technique of information extraction and incorporates a number of processing modules aimed at providing accurate marking without undue penalty for poor spelling and grammar. The question authors use an interface to the authoring tool which enables mark schemes to be represented as a series of templates.

Accurate marking is possible for many different and sometimes quite complex student responses, taking account of word order when appropriate. So, in answer to the question shown in Figure 2, a response of 'because oil is less dense than water' can be distinguished from 'because water is less dense than oil'. Similarly, a negated form of a correct response will be marked as incorrect (so 'the forces are not balanced' is distinguished from 'the forces are balanced').

The photograph shows a layer of oil floating on top of a glass of water. Why does the oil float?

The oil floats because the water has a biger density.

Check

Your answer is correct.

The oil floats because its density is less than that of the water, where density=$\dfrac{mass}{volume}$.

See Section 4.2.1.

Next

Figure 2 A free-text question, illustrating the correct marking of a response which includes a spelling mistake.

A novel feature of our project has been the use of student responses to developmental versions of the questions, themselves delivered online, to improve the answer matching. Previous users of the IAT software[11] and similar products[18] have used student responses to paper-based questions, 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 human marker as opposed to a computer.

Importance has been placed on the provision of instantaneous targeted feedback. Since the questions are offered to students via OpenMark, students are allowed several attempts, as described above. The feedback for incomplete or incorrect answers (as shown in Figure 3) 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 3 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 module, have been authored and refined in the light of student responses. Writing the initial answer matching can be a relatively quick process (typically taking around an hour) but amending it in the light of student responses is much more time consuming, taking more than a day for some questions. However the outcome is questions that can be re-used many times, and which engage students in a more meaningful way than more conventional e-assessment tasks.

Evaluation: human-computer marking comparison
A batch of student responses to each of seven free-text questions were marked independently by six course tutors, by the computer system 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. The percentage of responses where there was *any* variation in marking ranged between 4.8% (for Question 1 'What does an object's velocity tell you that its speed does not?', where the word 'direction' was an adequate response) and 64.4% (for Question 13, a more open-ended question: '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?'). However 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. Further improvements have been made to the IAT answer matching since the human-computer marking comparison took place in June 2007, and in July 2008, the marking of a new batch of responses was found to be in agreement with the question author for between 97.5% (for Question 13) and 99.6% (for Question 1) of the responses.

Mitchell et al.[19] have dentified the difficulty of accurately marking responses which include both a correct and an incorrect answer as 'a potentially serious problem for free text analysis'. Contrary to e-assessment folklore, responses of this type do not originate from students trying to 'beat the system' (for example by answering 'It has direction. It does not have direction') but rather by genuine misunderstanding, as exemplified by the response 'direction and acceleration' in answer to Question 1.

The computer marked this response 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. Whilst 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. In Mitchell et al's words 'while the characteristics of the set of creditworthy responses may be increased iteratively, algorithms for recognising incorrect science may approach the infinite'.

In acknowledging that computer-based marking of free-text answers will never be perfect, the inherent inconsistency of human markers (where different markers mark the same response in a different way or where one marker marks the same response differently on different occasions) should not be underestimated. If course tutors can be relieved of the drudgery associated with marking relatively short and simple responses, time is freed for them to spend more productively, perhaps in supporting students in the light of misunderstandings highlighted by the e-assessment questions or in marking questions where the sophistication of human judgement is more appropriate.

Evaluation: student observation
Each batch of developmental questions offered to students was accompanied by a short online questionnaire, and responses to this questionnaire indicate that a large majority of students enjoyed answering the questions and found the feedback useful. In order to further investigate student reaction to the questions and their use of the feedback provided, six student volunteers, from the course on which the questions were based, were observed attempting a number of short answer question alongside more conventional OpenMark questions. The students were asked to 'think out loud' and their words and actions were video-recorded.

Five of the six students were observed to enter their answers as phrases rather than complete sentences. It is not clear whether they 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 box in which the answer was to be input so they felt there was no need to repeat words from the question in the first part of the answer. One student was observed to enter his answers in long and complete sentences, which we initially interpreted as evidence that he was putting in as many keywords as possible in an attempt to match the required ones. However the careful phrasing of his answers makes this explanation unlikely; this student started off by commenting that he was 'going to answer the questions in exactly the same way as for a tutor-marked assignment' and it appears that he was doing just that.

Students were also observed to use the feedback in different ways. Some read the feedback carefully, scrolling across the text and making comments like 'fair enough'; these students frequently went on to use the feedback to correct their answer. However, evidence that students do not always read written feedback carefully came from the few instances where the system marked an incorrect response as correct. Students were observed to read the question author's answer (which appears when the student answer is either deemed to be correct or when it has been incorrect for three consecutive attempts) but not to appreciate that the response they had given was at variance with this. Being told that an incorrect answer is correct may act to reinforce a previous misunderstanding. Given the high accuracy of the computer's marking, this is not a common problem but it is an important one, as it is when a human marker fails to correct a student error.

Future developments
Modified versions of some of the questions developed have been incorporated, along with conventional OpenMark questions, into regular interactive computer marked assignments (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 relevant and instantaneous feedback and an opportunity for students to act on that feedback immediately.

Further information about this project and some sample questions are available on the author's COLMSCT website[20].

Acknowledgements

References

1. See for example, 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.
2. Yorke, M. (2003) Formative assessment in Higher Education: moves towards theory and the enhancement of pedagogic practice. *Higher Education*, **45**, 477-501.
3. Scouller, K.M. and Prosser, M. (1994) Students' experiences in studying for multiple choice question examinations, *Studies in Higher Education*, **19**(3), 267-279.
4. Broadfoot, P. (2008) Assessment for learners: assessment literacy and the development of learning power. In Havnes, Antin and McDowell, Liz (eds). Balancing dilemmas in assessment and learning in contemporary education, London: Routledge/Taylor and Francis, 213-224
5. Black, P. and William, D. (1998) Assessment and classroom learning, *Assessment in Education*, **5**(1), 7-74.
6. Gibbs, G and Simpson, C. (2004) Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education*, **1**, 3-31
7. 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.
8. Ramaprasad, A. (1983) On the definition of feedback, *Behavioral Science*, **28**, 4-13.
9. Sadler, D.R. (1989) Formative assessment and the design of instructional systems, *Instructional Science*, **18**,119-144.
10. Whitelock, D and Brasher, A (2006) *Roadmap for e-assessment*. Joint Information Systems Committee Report. http://www.jisc.ac.uk/elp_assessment.html (accessed 14th July 2008)
11. Mitchell, T., Aldridge, N., Williamson, W. and Broomhead, P. (2003). Computer based testing of medical knowledge. 7th International CAA Conference, Loughborough, UK. http://www.caaconference.com/pastconferences/2003/procedings/mitchell.pdf (accessed 14th July 2008)
12. Swithenby, S.J. (2006) Screen-based assessment. *New Directions*, **2**, 23-28
13. The COLMSCT website is at http://www.open.ac.uk/colmsct/
14. http://moodle.org/
15. Butcher, P.G. (in press) Online assessment at the Open University using open source software: Moodle, OpenMark and more. 12th International CAA Conference, Loughborough, UK, July 2008.
16. The OpenMark examples website is at http://www.open.ac.uk/openmarkexamples/index.shtml
17. 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.
18. Sukkarieh, J.Z, Pulman, S.G. and Raikes, N. (2003) Auto-marking: using computational linguistics to score short, free-text responses. 29thth International Association for Educational Assessment (IAEA) Annual Conference, Manchester
19. Mitchell, T., Russell, T., Broomhead, P. and Aldridge, N. (2002) Towards robust computerised marking of free-text responses. 6th International CAA Conference, Loughborough, UK. http://www.caaconference.com/pastconferences/2002/proceedings/Mitchell_t1.pdf (accessed 14th July 2008)
20. http://www.open.ac.uk/colmsct/projects/sallyjordan