



## Open Research Online

### Citation

Jordan, Sally and Mitchell, Tom (2009). e-Assessment for learning? The potential of short-answer free-text questions with tailored feedback. *British Journal of Educational Technology*, 40(2) pp. 371–385.

### URL

<https://oro.open.ac.uk/15270/>

### License

None Specified

### Policy

This document has been downloaded from Open Research Online, The Open University's repository of research publications. This version is being made available in accordance with Open Research Online policies available from [Open Research Online \(ORO\) Policies](#)

### Versions

If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding

**Paper accepted for publication in BJET special issue on assessment.**

***E-assessment for learning? The potential of short-answer free-text questions with tailored feedback***

**Sally Jordan and Tom Mitchell**

*Sally Jordan is a Staff Tutor in Science at the Open University in the East of England and a teaching fellow in the Centre for the Open Learning of Mathematics, Science, Computing and Technology (COLMSCT) and the Physics Innovations Centre for Excellence in Teaching and Learning (piCETL). She is one of several COLMSCT fellows who are developing and evaluating innovative e-assessment products, with the aim of maximising the learning function of interactive computer-marked assessment. Tom Mitchell is the founder of Intelligent Assessment Technologies, and is the software architect responsible for designing the company's FreeText and ExamOnline eAssessment products. Address for correspondence: Sally Jordan, COLMSCT, The Open University, Milton Keynes, MK7 6AA, UK. Email: s.e.jordan@open.ac.uk*

**Abstract**

A natural language based system has been used to author and mark short-answer free-text assessment tasks. Students attempt the questions online and are given tailored and relatively detailed feedback on incorrect and incomplete responses, and have the opportunity to repeat the task immediately so as to learn from the feedback provided. The answer matching has been developed in the light of student responses to the questions. A small number of the questions are now in low-stakes summative use, alongside other e-assessment tasks and tutor-marked assignments, to give students instantaneous feedback on constructed response items, to help them to monitor their progress and to encourage dialogue with their tutor. The answer matching has been demonstrated to be of similar or greater accuracy than specialist human markers. 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. We discuss features of appropriate items for assessment of this type.

**Introduction**

E-assessment enables feedback to be delivered instantaneously. This provides an opportunity for students to take immediate action to 'close the gap' between their current level and a reference point, and thus for the feedback to be effective (Ramaprasad, 1983; Sadler, 1989). However concern has been expressed that conventional e-assessment tasks can encourage a surface approach to learning (Scouller & Prosser, 1994; Gibbs, 2006).

Assessment items can be broadly classified as selected response (for example multiple-choice) or constructed response (for example short-answer). Short-answer constructed response items require the respondent to construct a response in natural language and to do so without the benefit of any prompts in the question. This implies a different form of cognitive processing and memory retrieval when compared with

selected response items (Nicol, 2007). Short-answer constructed response items are highly valued in traditional paper-based assessment and learning, but have been almost completely absent from computer-based assessment due to limitations and perceived limitations in computerised marking technology.

Recent developments have seen the introduction of natural language based assessment engines. One such engine, developed by Intelligent Assessment Technologies (IAT), has been deployed by the UK Open University (OU) to support the learning of adult distance learners.

### **Software for the marking of free-text answers**

Perhaps the most well-known system for the e-assessment of free text is e-rater (Attali & Burstein, 2006), an automatic essay scoring system employing a holistic scoring approach. The system is able to correlate human reader scores with automatically extracted linguistic features, and provide an agreement rate of over 97% for domains where grading is concerned more with writing style than with content. A different technique which shows high promise is that of Latent Semantic Analysis (LSA) (Landauer, Foltz & Laham, 2003). LSA has been applied to essay grading, and high agreement levels obtained. These techniques are more suited to marking essays than short-answer questions, since they focus on metrics which broadly correlate with writing style, augmented with aggregate measures of vocabulary usage. Computerised marking of short-answer questions on the other hand, is concerned with marking for content above all else.

C-rater is a short-answer marking engine developed by Education Testing Service (ETS) (Leacock & Chodorow, 2003). The system represents correct (i.e. model) answers using 'canonical representations', which attempt to represent the knowledge contained within an answer, normalised for syntactic variations, pronoun references, morphological variations, and the use of synonyms. Reported agreement with human markers is of the order of 85%.

In the UK, Pulman and Sukkarieh (2005) used information extraction techniques for marking short-answer questions. Their system can be configured in a 'knowledge engineering' mode, where the information extraction patterns are discovered by a human expert, and a 'machine learning' mode, where the patterns are learned by the software. The 'knowledge engineering' approach is more accurate and requires less training data but it requires considerable skill and time of a knowledge engineer.

The software developed by Intelligent Assessment Technologies and used by the Open University is most closely related to the system developed by Pulman and Sukkarieh, in that it borrows from information extraction techniques. The main strength of the IAT system is that it provides an authoring tool which enables a question author with no knowledge of natural language processing (NLP) to use the software.

### **Intelligent Assessment Technologies' software**

IAT's free-text assessment software comprises the marking engine itself, provided as a web service, and an authoring tool, used to configure the marking rules for each question.

### *The free-text marking engine*

The IAT software employs NLP tools and techniques to provide computerised marking of short free-text answers and is particularly suited to marking questions requiring an answer of one or two sentences. The software includes a number of modules developed to enable accurate marking without undue penalty for errors in spelling, grammar or punctuation.

The marking engine performs a match between free-text student responses and predefined computerised model answers, in an analogous process to that used by human markers when marking free-text answers for content. The model answers are represented as syntactic-semantic templates, each specifying one particular form of acceptable or unacceptable answer. Figure 1 shows a template for the model answer ‘The Earth rotates around the Sun’. This template will match a student response if that response contains one of the stated verbs (*rotate, revolve, orbit, travel, move*) with one of the stated nouns (*earth, world*) as its subject, and *around/round the sun* in its preposition. Verbs in the student response are lemmatised (i.e., reduced to their base form).

*Figure 1: Template for the model answer ‘The Earth rotates around the Sun’*

The following responses would all be matched by the template shown in Figure 1:

- The World rotates around the Sun.
- The Earth is orbiting around the Sun.
- The Earth travels in space around the Sun.

However, incorrect responses such as ‘The Sun orbits the Earth’ would not be matched by the template, a significant improvement over simple ‘bag of words’ keyword-matching systems. Similarly, negated forms of responses can be recognised as such without specific action from the question author, so an answer of ‘The forces are not balanced.’ can be distinguished from ‘The forces are balanced.’

The development of the templates in a computerised mark scheme is an offline process, achieved using IAT’s FreeText Author software. Once the mark scheme for a question has been developed, it can be used by the engine to mark student answers. Incoming free-text answers are processed by a sentence analyser, and the output is matched against each mark scheme template. The result of the matching process determines the mark awarded to the response. In addition, appropriate feedback can be associated with each model answer.

### *The FreeText Author user interface*

The FreeText Author user interface was designed to simplify the task of generating templates by shielding the user from the complexities of NLP, allowing them to concentrate on tasks such as identifying model answers for the question and the keywords for each model answer. FreeText Author’s point-and-click user interface (Figure 2) is designed for use by subject experts (lecturers and examiners) rather than NLP experts.

*Figure 2: The FreeText Author user interface*

FreeText Author's main components are a mark scheme panel, a model answer list, a synonym editor and a response list. The mark scheme panel includes lists of acceptable and unacceptable top-level mark scheme answers, mimicking the layout of a traditional paper-based mark scheme. Associated with each mark scheme answer is a number of model answers, each representing specific acceptable phrasings for the corresponding mark scheme answer. Templates are generated automatically from the model answers, using a machine learning algorithm included in FreeText Author.

When a user enters a new model answer they must specify the keywords (words which must be found in a student response before it is even possible that the model answer will be matched). The synonym editor allows the user to add synonyms for each keyword, with suggestions provided from an inbuilt thesaurus and from an analysis of other model answers and the list of responses.

The development of a computerised mark scheme involves an iterative process of adding and modifying model answers. At each iteration, the efficacy of the change can be tested by applying the mark scheme to the response list. Responses can be added to the response list manually (by typing in new responses) or by importing responses from an external file (as may be extracted from a database of student responses acquired when the question is trialled or used in a live assessment).

### **Use of FreeText Author at the Open University**

The Open University was set up in 1969 to provide degree programmes by distance learning. There are no formal entrance requirements for undergraduate courses. Students are usually adults and most are in employment and/or have other responsibilities, so study part-time. Most students are allocated to a personal tutor, but opportunities for face-to-face contact are limited, and although support by telephone is encouraged and increasing use is made of email and online forums, there remain a few students for whom communication with tutors is particularly difficult. Extensive use is made of tutor-marked assignments and tutors are encouraged to return these quickly, but in the lonely world of the distance learner, instantaneous feedback on online assessment tasks provides a way of simulating for the student 'a tutor at their elbow' (Ross, Jordan & Butcher, 2006, p.125). 'Little and often' assignments can be incorporated at regular intervals throughout the course, assisting students to allocate appropriate amounts of time and effort to the most important aspects of the course (Gibbs & Simpson, 2004). Finally, the Open University is making increasing use of e-learning, so e-assessment is a natural partner (Mackenzie, 2003), providing alignment of teaching and assessment modes (Gipps, 2005).

In order to provide Open University students with useful instantaneous feedback, the IAT short-answer questions are embedded within the OpenMark assessment system, which was developed by the Open University but is now open source. OpenMark provides a range of other question types allowing for the free-text entry of numbers, scientific units, simple algebraic expressions and single words as well as drag-and-drop, hotspot, multiple-choice and multiple-response questions. However the significant feature for the current project is OpenMark's ability to provide students with multiple attempts at each question, with the amount of feedback increasing at each attempt. If the questions are used summatively, the mark awarded decreases after each attempt, but the presence of multiple attempts with increasing feedback remains a feature. Thus, even in summative use, the focus is on assessment for learning. At the

first attempt an incorrect response will result in very brief feedback, designed to give the student the opportunity to correct their answer with the minimum of assistance. If the student's response is still incorrect or incomplete at the second attempt, they will receive a more detailed hint, wherever possible tailored to the misunderstanding which has led to the error and with a reference to the course material. After a third unsuccessful attempt, or whenever a correct answer has been given, the student will receive a model answer. The three feedback stages are illustrated in Figure 3.

*Figure 3: Increasing feedback received on three attempts at an IAT short-answer question embedded within OpenMark*

The OpenMark e-assessment system sits within the Moodle virtual learning environment (Butcher, 2008). The Moodle Gradebook enables students to monitor their own progress, encouraging sustainable self-assessment practices (Boud, 2000). The tutor's view of the Gradebook encourages dialogue between student and tutor (Nicol & Milligan, 2006).

The range of possible correct and incorrect answers to each free-text short-answer question means that it is important to develop answer matching in the light of responses gathered from students at a similar level to those for whom the questions are intended. Previous users of the IAT software and similar products have used student responses to paper-based questions (Mitchell, Aldridge, Williamson & Broomhead, 2003; Sukkariah, Pulman & Raikes, 2003) but this approach assumes that there are no characteristic differences between 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. In the current project, student responses to the online developmental versions of the questions have been used to improve the answer matching.

The IAT software sets a flag if a student response fails to match a model answer but is recognised as being close to one (for example, if keywords are matched but the marking engine is not able to parse the student response). When questions are marked in batch mode, as described in Mitchell et al. (2003), these flags draw attention to responses requiring intervention from a human-marker, and/or indicate answer matching requiring further refinement. In the current project, where marking is online, it was initially decided to give students the benefit of the doubt for all such responses.

A second system of flags was incorporated into FreeText Author by IAT specifically to support the OU assessment model, enabling tailored feedback to be provided for incorrect and incomplete responses. The feedback is written by the question author, in the light of information about common misconceptions gathered from previous student responses. The answer matching for feedback operates separately from the answer matching for grading, which enables detailed and personalised feedback to be given to students without compromising marking accuracy.

Seventy-eight short-answer questions with tailored feedback have been authored by a full-time member of the Open University's lecturing staff (the first-named author) and a part-time associate lecturer. Neither is trained in either NLP or computer programming. These questions were offered to students on three presentations of an introductory, interdisciplinary science course, as a formative-only add-on, and answer

matching was improved during the first presentation prior to evaluation (briefly described below) during the second presentation. Further improvements were made during the third presentation and a small number of the questions are now in low-stakes summative use in regular computer-marked assignments (iCMAs) alongside more conventional OpenMark questions.

## **Evaluation**

### *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', both the computer's and each course tutor's marking of each response to each question were compared against:

- The median of all the course tutor's marks for that response;
- The marking of individual responses by the author of the questions. This marking was done 'blind', without knowledge of the way in which the course tutors had marked the question or the way in which the IAT system had responded to each particular response. However the author was very familiar with the mark schemes and model answers that the IAT system was applying;
- The raw IAT score, without credit being given for answers that were flagged by IAT as being close to a known model answer;
- The score for the individual response as delivered by OpenMark, which was the IAT score amended to give credit when a response is flagged as being 'close' to a known model answer.

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 four of the questions, the marking of all the markers (including the computer system, with and without adjustment for flagging) was indistinguishable at the 1% level (see Table 1). For the other three questions, the markers were marking in a way that was significantly different. However in all cases, the mean mark allocated by the computer system (again, with or without adjustment for flagging) 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 A '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 G, 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 for each of the questions, the majority of the variation was caused by discrepancies in the marking of the course tutors. On some occasions one human marker consistently marked in a way that was different from the others; on other occasions an individual marked inconsistently (marking a response as correct, when an identical one had previously been marked as incorrect, or vice versa). Divergence of human marking could frequently be attributed to insufficient detail in the marking guidelines or to uncertainty over whether to give credit for a partially

correct solution. However, there were also some errors caused by slips and by poor subject knowledge or understanding.

*Table 1: Some data from the human-computer marking comparison*

For six of the questions, the marking of the computer system was in agreement with that of the question author for more than 94.7% of the responses (rising as high as 99.5% for Question A). For Question G, the least well developed of the questions at the time the comparison took place, there was agreement with the question author for 89.4% of the responses. Further improvements have been made to the 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 G) and 99.6% (for Question A) of the responses. This is in line with a previous study of the IAT engine's marking (Mitchell et al., 2003) where an accuracy of >99% was found for simple test items.

Mitchell et al. (2002) have identified 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 A. 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 FreeText Author by the addition of a 'do not accept' mark-scheme) it is not realistic to make provision for all flawed answers of this type.

For two of the questions in the human-computer marking comparison, the combined IAT/OpenMark marking was found to be more accurate if credit was only given for answers that exactly matched a model answer (i.e., not if they were flagged by the IAT marking engine as being close to one). This can be explained by the fact that if the correct keywords are given but in the incorrect order (for example, 'gravitational energy is converted to kinetic energy' instead of 'kinetic energy is converted to gravitational energy') the IAT marking engine accurately marks the response as incorrect but sometimes flags the incorrect response as being close to the correct model answer. The adjustment for flagging is now only applied in questions where it is known not to cause problems of this sort.

*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 was 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.

### **Using the authoring tool: what makes a good question?**

One of the barriers to wider take-up of e-assessment is the difficulty or perceived difficulty of writing appropriate assessment tasks, with the inevitable knock-on effect on costs. Even when the driver for development is improvement to the student learning experience, as Gipps (2005, p. 178) says 'those who are driving online teaching, learning and assessment in higher education cannot ignore the resource issues'. For this reason, the ease of use of the authoring tool by inexperienced users was monitored carefully.

After the initial training phase, a question author experienced in writing a range of assessment tasks (including more conventional e-assessment questions) was able to write short-answer free-text questions and appropriate answer matching with relative ease. The time spent in the initial writing of the question and answer matching varied between a few minutes and several hours, depending on the complexity of the question. Amending the question and the answer matching in the light of student responses was even more dependent on the complexity of the question, taking more than a day for some questions. However the accuracy of the answer matching was undoubtedly improved by its development in the light of real student answers. It is also worth noting that the questions deployed by the OU tended to be at the more complex end of the short-answer spectrum, and that the inclusion of detailed feedback for complex mark schemes added to the time required for development. By comparison, questions deployed to test medical knowledge in other institutions have been developed and moderated in minutes rather than hours (Mitchell et al., 2003). The rule of thumb is that it is possible to develop more open and complex questions,

but that more time is required to do so. Within the Open University, as reported more generally (e.g. Sim, Holifield & Brown, 2004; Conole & Warburton, 2005), the greatest barrier to wider take up of rich assessment tasks of this type appears to be the time required to learn how to use novel software and to develop high-quality questions; to some extent this can be ameliorated by the provision of appropriate staff development.

Of the 78 questions originally authored, four were deemed to be unworkable and removed during development. In a further 13 cases, changes were made to the wording of the questions themselves because it appeared that students had been confused by the original question or it transpired that the responses generated were too difficult to match. In most cases the changes of wording were minimal, but occasionally they acted to more tightly constrain the student responses. So ‘You are handed a rock specimen that consists of interlocking crystals. How would you decide, from its appearance, whether it is an igneous or a metamorphic rock?’ became ‘You are handed a rock specimen that consists of interlocking crystals. How could you be sure, from its appearance, that this was a metamorphic rock?’ The second version of the question, although tightly constrained and answerable in a very short sentence, assesses more than recall – students are required to apply knowledge from the course to a new scenario.

*Figure 4: Accurate marking of a relatively complex response*

Questions are likely to be suitable for computerised marking using the IAT marking engine if correct answers can be given in short phrases or simple sentences and the difference between correct and incorrect answers is clear-cut. Questions are likely to be unsuitable if there are too many ways in which a correct response can be expressed or if responses are complex in nature. Reliable answer matching has been obtained for the question shown in Figure 4, where a correct answer must mention that the rock is formed from magma [molten rock], that the magma has cooled [or crystallised/crystallized/solidified] and that the cooling has taken place slowly [or over a long period of time/deep underground]. However, if students are required to write about two or more separate concepts in one answer, matching can be difficult. At present the most reliable solution to this problem is to split the question into separate parts. Sometimes this can be achieved without severe detriment to the assessment task, as shown in Figure 5. In other cases, splitting the task would substantially alter its function and so is unlikely to be a desirable way forward. In addition, in more discursive disciplines, questions are less likely to have clear ‘right’ and ‘wrong’ answers.

*Figure 5: A question with separate answer matching for two parts*

### **Reflection**

Interactive computer-marked assignments are being used increasingly at the Open University, primarily for their capacity to encourage student engagement and to deliver instantaneous teaching feedback; some of these iCMAs now include short-answer free-text questions. However, in almost all cases, the iCMAs exist alongside more conventional tutor-marked assignments as part of an integrated assessment strategy, and take-up is greater in some faculties (especially Science; Mathematics,

Computing and Technology and Health and Social Care) than others (especially Social Science and Arts).

A computerised system has been shown to accurately mark short-answer free-text questions and to deliver tailored feedback on incorrect and incomplete responses. Accuracy of marking is important even in formative-only use, to ensure that students receive the correct feedback. While acknowledging that computerised marking of free-text answers will never be perfect, the inherent inconsistency of human markers should not be underestimated; computerised marking is inherently consistent (Conole & Warburton, 2005). 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.

Short-answer questions have been delivered to students alongside more conventional e-assessment tasks, and early indications are that they can be as reliable as selected response items (where students may have guessed, as discussed by Sim et al., 2005, or have arrived at their answer by working backwards from the options provided). It is possible for a human to monitor exactly what the student has entered, not just the option that the student has selected (without any indication as to why that option has been chosen). Further work is planned to investigate the effect of low-stakes summative use of e-assessment and to further investigate how students react to feedback from a computer.

### **Acknowledgements**

The financial support of the Centre for Open Learning of Mathematics, Science, Computing and Technology (COLMSCT) is gratefully acknowledged, as is assistance from many colleagues, especially Barbara Brockbank, Philip Butcher, Laura Hills and Stephen Swithenby (COLMSCT) and Tim Hunt (Learning and Teaching Solutions, the Open University).

### **References**

Attali, Y. & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning and Assessment*, 4, 3.

Butcher, P. (2008). Online assessment at the Open University using open source software: Moodle, OpenMark and more. 12<sup>th</sup> International CAA Conference, Loughborough, UK. Retrieved 30 October 2008 from <http://www.caaconference.com/pastConferences/2008/proceedings/index.asp>

Conole, G & Warburton, B. (2005) A review of computer-assisted assessment. *ALT-J*, 13, 1, 17–31.

Gibbs, G. & Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education*, 1, 3–31.

Gibbs, G. (2006). Why assessment is changing. In C. Bryan & K.V. Clegg, K.V. (Eds), *Innovative assessment in higher education* (pp. 11–22). London: Routledge.

Gipps, C.V. (2005). What is the role for ICT-based assessment in universities? *Studies in Higher Education*, 30, 2, 171–180.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds), *Automated essay scoring: a cross-disciplinary perspective* (pp. 87–112). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Leacock, C. & Chodorow, M. (2003). C-rater: Automated Scoring of Short-Answer Questions. *Computers and Humanities*, 37, 4, 389–405.

Mackenzie, D. (2003). Assessment for e-learning : what are the features of an ideal e-assessment system?. 7<sup>th</sup> International CAA Conference, Loughborough, UK. Retrieved 30 October 2008 from <http://www.caaconference.com/pastConferences/2003/proceedings/index.asp>

Mitchell, T., Russell, T., Broomhead, P. & Aldridge, N. (2002). Towards robust computerised marking of free-text responses. 6<sup>th</sup> International CAA Conference, Loughborough, UK. Retrieved 30 October 2008 from <http://www.caaconference.com/pastConferences/2002/proceedings/index.asp>

Mitchell, T., Aldridge, N., Williamson, W. & Broomhead, P. (2003). Computer based testing of medical knowledge. 7<sup>th</sup> International CAA Conference, Loughborough, UK. Retrieved 30 October 2008 from <http://www.caaconference.com/pastConferences/2003/proceedings/index.asp>

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. & Milligan, C. (2006). Rethinking technology supported assessment practices in relation to the seven principles of good feedback practice. In C. Bryan & K.V. Clegg, K.V. (Eds), *Innovative assessment in higher education* (pp. 64–77). London: Routledge.

Pulman, S. & Sukkarieh, J. (2005) Automatic Short Answer Marking. Proceedings of the 2nd Workshop on Building Educational Applications Using NLP, Ann Arbor, June 2005. Retrieved 30 October 2008 from <http://www.comlab.oxford.ac.uk/people/publications/date/Stephen.Pulman.html>

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science*, 28, 4–13.

Ross, S.M., Jordan, S.E & Butcher, P.G. (2006). Online instantaneous and targeted feedback for remote learners. In C. Bryan & K.V. Clegg, K.V. (Eds), *Innovative assessment in higher education* (pp. 123–131). London: Routledge.

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

Scouller, K.M. & Prosser, M. (1994). Students' experiences of studying for multiple choice question examinations. *Studies in Higher Education*, 19, 3, 267–279.

Sim, G., Holifield, P. & Brown, M. Implementation of computer assisted assessment: lessons from the literature. *ALT-J*, 12, 3, 215–229.

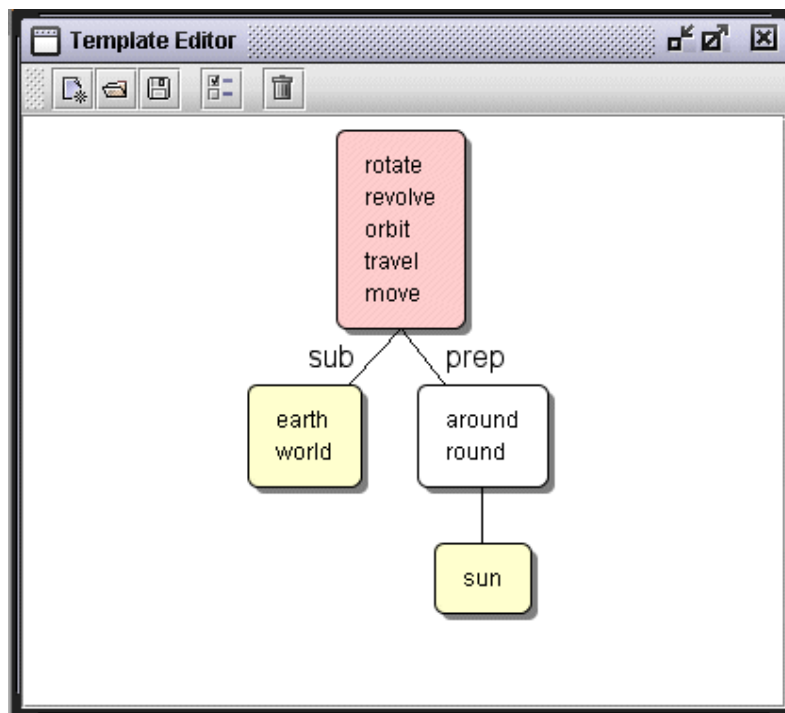
Sukkarieh, J.Z, Pulman, S.G. & Raikes, N. (2003). Auto-marking: using computational linguistics to score short, free-text responses. 29<sup>th</sup> International Association for Educational Assessment (IAEA) Annual Conference, Manchester, UK. Retrieved 30 October 2008 from <http://www.comlab.oxford.ac.uk/people/publications/date/Stephen.Pulman.html>

## Tables and Figures

*Table 1: Some data from the human-computer marking comparison*

<i>Question</i>	<i>Number of responses in analysis</i>	<i>Probability that marking of 6 human markers and computer is equivalent</i>	<i>Percentage of responses where computer marking was in agreement with question author</i>
Question A	189	$p > 0.99$	99.5%
Question B	246	$p < 0.01$	96.3%*
Question C	150	$p < 0.001$	94.7%
Question D	129	$p > 0.96$	96.9%*
Question E	92	$p > 0.71$	98.9%
Question F	129	$p > 0.15$	97.7%
Question G	132	$p < 0.0001$	89.4%

\* All results shown are for marking with credit for responses that are close to a model answer without matching it exactly. For Questions B and D, the marking was more accurate without this adjustment, with 97.6% agreement with the question author in each case.



*Figure 1: Template for the model answer 'The Earth rotates around the Sun'*

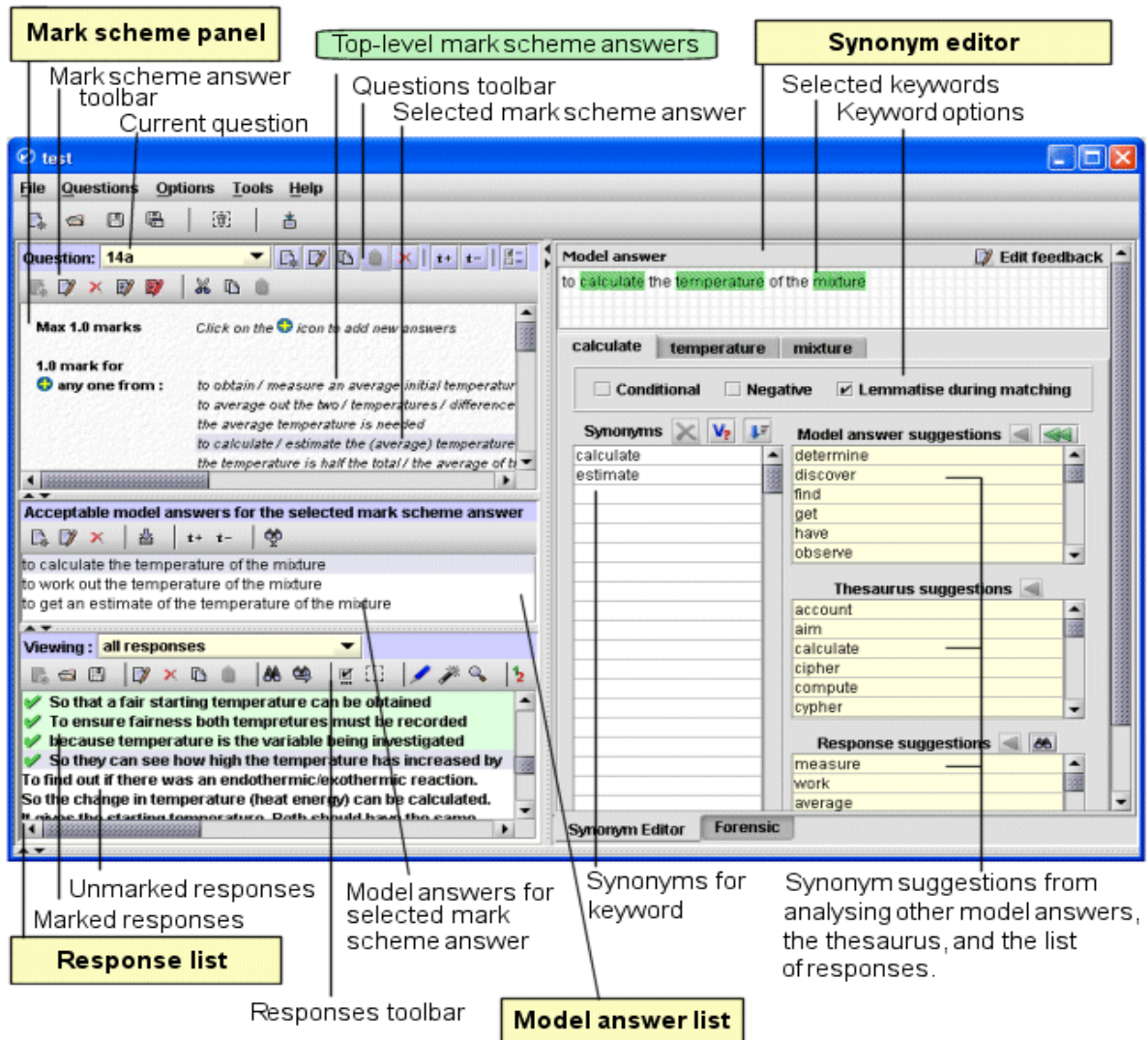



Figure 2: The FreeText Author user interface

<p>If two particles which are 4 metres apart are moved to a new separation of 1 metre, what happens to the gravitational force between them? Be as specific as possible.</p> <p>As the separation decreases, the gravitational force increases.</p>	<p>Your answer appears to be incorrect or incomplete in some way.</p> <p>Have another go, remembering to express your answer as a simple sentence.</p> <p>Try again</p>
<p>Check</p> <p>If two particles which are 4 metres apart are moved to a new separation of 1 metre, what happens to the gravitational force between them? Be as specific as possible.</p> <p>It will be four times bigger.</p>	<p>Your answer still does not appear to be correct.</p> <p>You are correct to say that the force increases, but you are not correct to say that it increases by a factor of four. Newton's law of gravity states that the gravitational force between two particles is inversely proportional to the square of their separation (see Block 11 Section 8.1). So when the separation is decreased by a factor of 4, what happens to the gravitational force between the particles?</p> <p>Try again</p>
<p>Check</p> <p>If two particles which are 4 metres apart are moved to a new separation of 1 metre, what happens to the gravitational force between them? Be as specific as possible.</p> <p>The inverse square law means that it will be sixteen times bigger.</p> <p>Check</p>	<p>Your answer is correct.</p> <p>Newton's law of gravity states that the gravitational force between two particles is inversely proportional to the square of their separation. So when the distance between the two particles is decreased by a factor of 4, the gravitational force is increased by a factor of 16.</p> <p><input type="checkbox"/> If you believe that the computer has marked your answer inaccurately please tick this box and your answer will be reviewed by a tutor.</p> <p>Next</p>

Figure 3: Increasing feedback received on three attempts at an IAT short-answer question embedded within OpenMark



The photograph shows an outcrop of granite near Land's End in Cornwall (UK). How is an igneous rock with large crystals (such as this granite) formed?



They are formed by the slow chrystallization of molton rock (magma) deep under the surface of the Earth.

Check

Your answer is correct.

Igneous rocks are formed from molten rock (magma) which has cooled and solidified. In the case of granite, this cooling will have happened very slowly deep underneath the Earth's surface. The granite will only have been exposed at the Earth's surface after overlying rocks have been removed by erosion.

Figure 4: Accurate marking of a relatively complex response

Light is shone on a metal and photoelectrons are emitted. The frequency of the light is then reduced. What does the photoelectric effect tell you about the two changes that might occur? Write one possible outcome in each of the boxes provided.

First

There might be no photoelectrons emitted at all.

Second

The maximum kinetic energy of the photoelectrons that are emitted might be less.

Check

Your answer is correct.

Decreasing the frequency of the light shining on the metal will decrease the maximum kinetic energy of the emitted photoelectrons (the opposite situation is illustrated in Block 7 Figure 9.1(b)). In addition, there is a lower limit to the frequency of the radiation, beyond which no photoelectrons are emitted (see Block 7 Figure 9.1(c)).

If you believe that the computer has marked your answer inaccurately please tick this box and your answer will be reviewed by a tutor.

Next

Figure 5: A question with separate answer matching for two parts