

Short answer marking engines



Short answer marking engines

What are short answers?

“Short answers” implies free text entry, requiring answers that have to be constructed rather than selected, ranging from phrases to (rarely) 3 to 4 sentences. In the context of e-assessment, “short text marking” is an abbreviation of the phrase “automated marking of short free text responses to test items”. Candidates making a “free text response” are required to provide an answer to a test item in their own words as opposed to making a choice from a list of possible answers. There is no absolute definition of “short” but it is usually taken to mean that the marking algorithm can attempt a syntactic analysis of the response followed by some form of semantic analysis, something that is beyond the state of the art above 20 words or so.

Why use short answers in assessments?

Multiple choice items (in all their variety) are a very efficient means of assessment, particularly though not exclusively, at lower taxonomic levels (eg knowledge, understanding and recall). Their ability to be fully computerised in delivery, marking and analysis make them popular and their use is especially associated with educational testing in the United States. Good multiple choice items are, however, notoriously hard to construct and have been criticised because they test the candidate’s ability to select an answer rather than to freely construct one. Short answer questions demanding short (free) text responses can be more testing in that candidates have to create their own response rather than choosing amongst plausible alternatives.

An assessment regime should use a variety of question types and modes of response if it is going to adequately test a candidate’s performance and short text marking engines can increase this variety as they allow the inclusion of short answer questions which can also be effectively computerised.

Question types that marking engines are good at marking

Short answer marking engines work best with questions producing convergent answers, that is, where there are a limited (though large) set of answers that the examiner is looking for, where knowledge and understanding is being tested, and where content is important rather than style.
http://www.iaea2008.cambridgeassessment.org.uk/ca/digitalAssets/164792_Siddiqi_Harrison.pdf. This might be seen as a limitation but most summative and much formative assessment is aimed at finding out what the student knows or can deduce. Questions leading to divergent answers are generally seeking to explore the quality of thought of the student and, as such, are not well marked by the marking engines (and indeed, do not generally elicit consistent marks from human assessors).

The testing of ability to construct a sustained argument, which is often tested in essay questions, is not well-addressed through short text assessment. The best-known program for marking essays is ETS¹ e-rater (ETS) which compares correlates of essay quality (eg style, vocabulary, length) with those of a battery of pre-graded scripts, finds the best match and scores accordingly. It is admirably suited to a tradition where scripts are routinely doubly marked by examiners but is not used as a lone marking system. There is no sense in which the program ‘understands’ the content of the essay.

Questions that marking engines find it hard to mark

Short text marking engines do not cope well with questions where there is an unpredictable range of acceptable answers, eg “What is democracy?” or where the answer is complex eg “Was Churchill a good prime minister and why?” Where marking engines have difficulty with scoring an item, closer moderation of the item will often reveal ambiguities or infelicities which can be corrected. Used routinely, this moderation process can suggest improvements to assessment and teaching.

1 Education Testing Services, based in Princeton and elsewhere

Feedback

Short answer marking engines can be used for assessments which are being used formatively, summatively or intended to achieve both outcomes. Used formatively, not only can the software provide an instant mark but it can also give answer-specific feedback. Several attempts at answering a question can be allowed (and recorded) and being a part of the learning process, cheating is pointless. Used summatively, the provision of feedback may be more problematic (but see the Open University case study below).

Common characteristics of how the marking engines work

Widely used short text marking engines include:

- C-rater (Educational Testing Services)
- systems developed by Jana Sukkarieh and Stephen Pulman at Oxford (later continued at Cambridge Assessment) and
- the marking engine and authoring tool developed by Intelligent Assessment Technologies (IAT).

These have all been developed over some years and offer HEIs the opportunity to replace or augment their e-assessment systems at relatively low cost.

The systems have been developed independently but

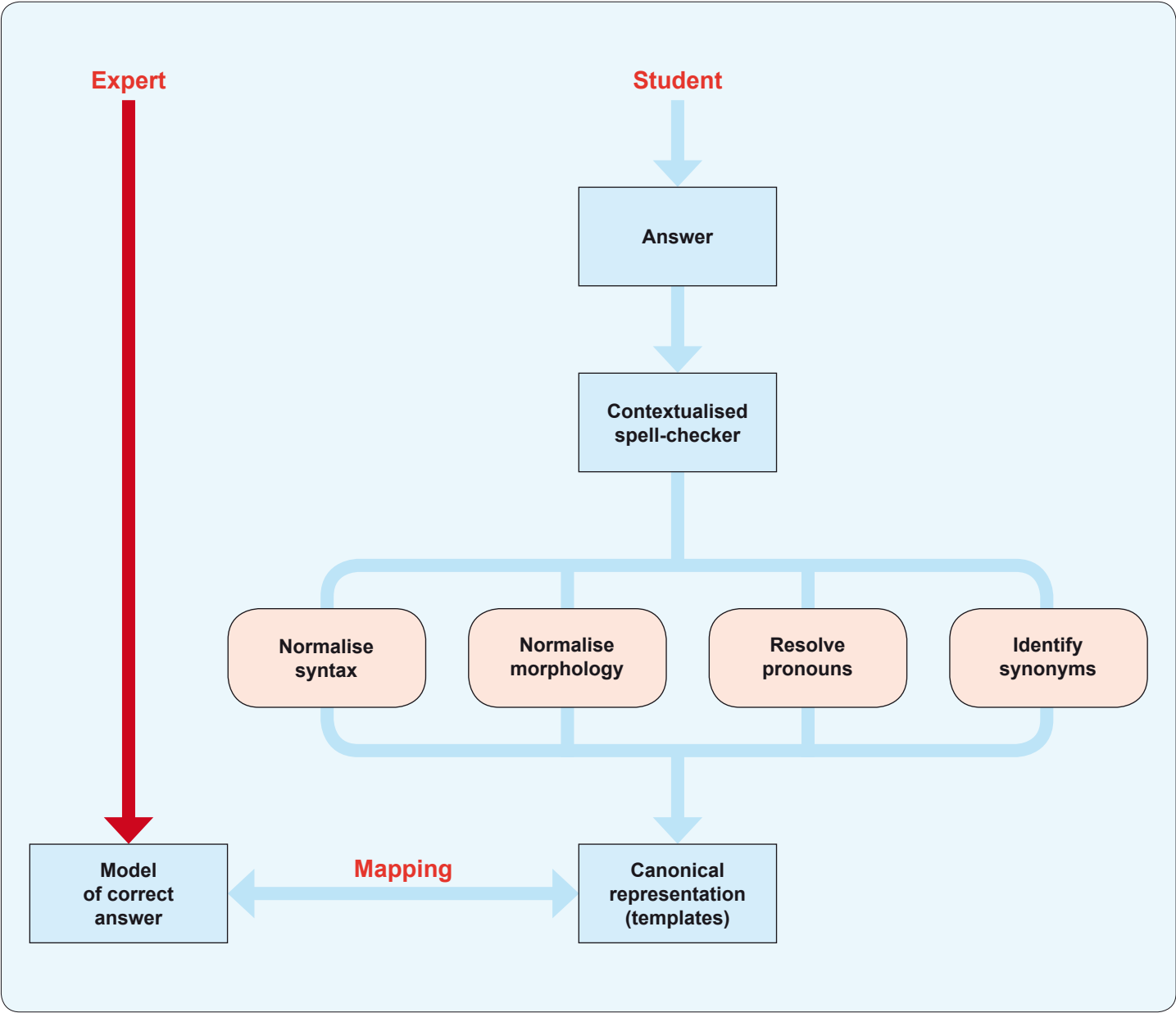
show marked convergence in how they work. All perform a limited analysis of student responses using broadly similar computational linguistics techniques. They follow the same basic pattern.

1. examiner generates top-level marking guideline – what the examiner is looking for
2. a small set of model answers (variously called ‘templates’, ‘canonical representations’ or ‘patterns’) is produced and developed using sample responses
3. student responses are analysed to see whether they are credit-worthy paraphrases of the model answers
4. there is a moderation process where the scoring is checked

There is generally human intervention at 1, 2 and 4 but, increasingly, tools are provided to make the generation of templates and moderation more straightforward and efficient. There are differences between (and within) systems as to how much of the pattern matching is done by the machine and how much by a human and for any set of results, there is (as one might expect) a trade off between the complexity of the answers, the amount of human input and the accuracy of the final scores. Pattern-matching can be set to be strict or lax, generating higher proportions of false negatives or false positives respectively, and systems can run with more or less human intervention.

Fully automated machine learning approaches have been attempted (eg see Sukkarieh, J. Z. and Pulman, S. G. (2005)).

Comparing the student response with the model answer



The specific model illustrated is C-rater. The expert is helped to produce model answers through the use of an application called Alchemist through which the expert stipulates what essential points are required in order to gain credit and identify likely synonyms. Students' responses are used to train the system (with human intervention).

- | | | |
|---|--------------------|---|
| 1 | spell-check | Program is forewarned of likely words and will preferentially correct to these. It cannot, however, correct misspellings which have resulted in a different English word eg 'umber'. |
| 2 | normalizing syntax | Program identifies verb, subject and object and reconfigures as ordered list (tuple) |
| 3 | morphology | Endings (inflections) are removed eg subtracts, subtracting, subtraction are all reduced to subtract. Negative prefixes are also stripped out and replaced with 'not' viz 'is unfair' to 'is not fair'. |
| 4 | pronoun resolution | Program identifies all the noun phrases that precede the pronoun and selects the one that the pronoun is most likely to refer to. |
| 5 | synonyms | Synonyms are generated automatically from a dictionary but synonyms are preferred which match the context of the rest of the answer. |

Links

Leacock, C and Chodorov, M. (2003). C-rater: Automated Scoring of short answer Questions. Computers and Humanities, 37, 4, 389-405. Also http://www.ets.org/Media/Research/pdf/erater_examens_leacock.pdf

Sukkarieh, J. Z. , Pulman, S. G. and Raikes, N. (2003) Auto-marking: using computational linguistics to score short free text responses. <http://www.clg.ox.ac.uk/pulman/pdppapers/AUTOMARKING2.htm>

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

<http://www.intelligentassessment.com>

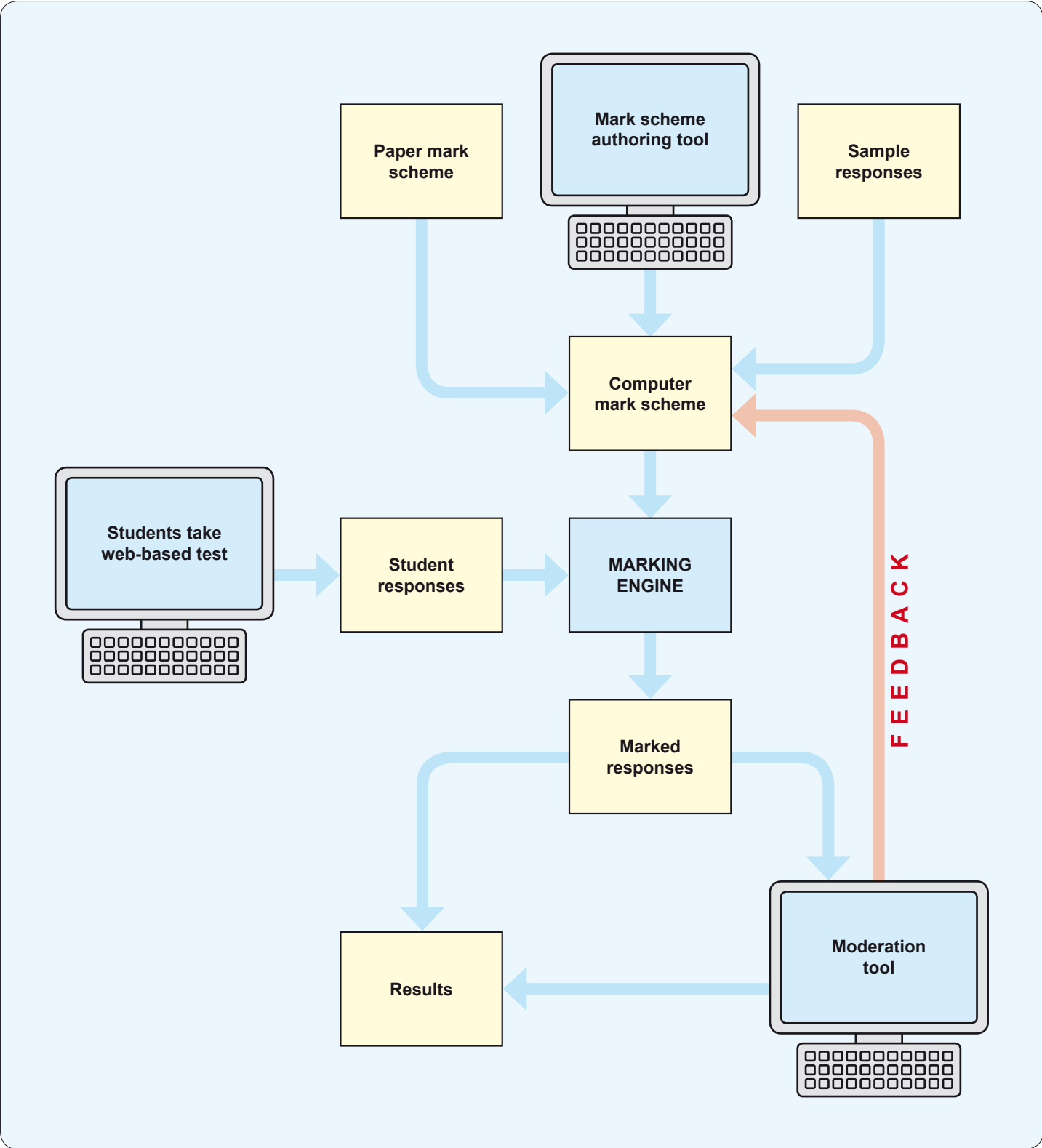
There is a comparative study at http://www.iaea2008.cambridgeassessment.org.uk/ca/digitalAssets/164792_Siddiqi_Harrison.pdf

The process in practice

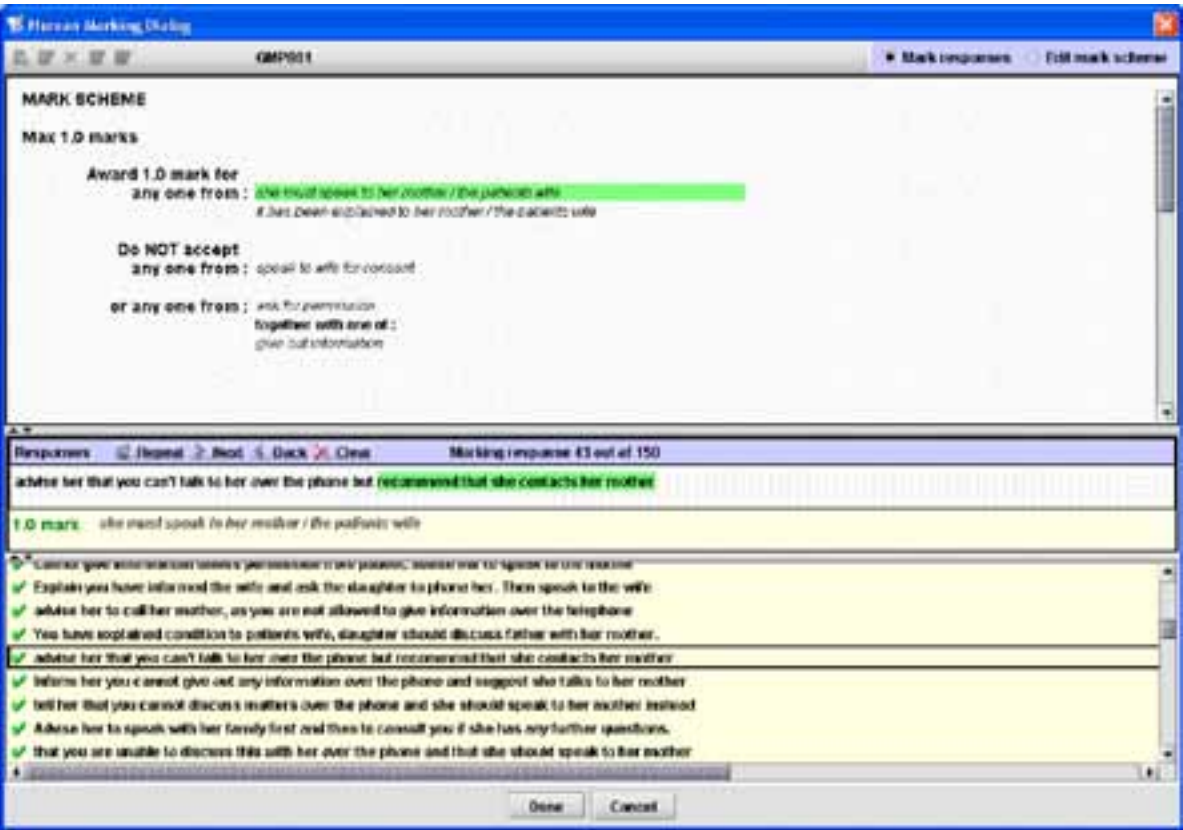
Intelligent Assessment Technology's system is used as an exemplar.

- 1. The examiner writes the item, possibly using an authoring tool or other preferred software.
- 2. The paper-based mark scheme, and any sample responses if available, are used to create computerised mark scheme files using the authoring tool. The templates may be tested against more sample answers. This process will often reveal improvements that can be made to questions. The items and their markschemes are collated into a test.

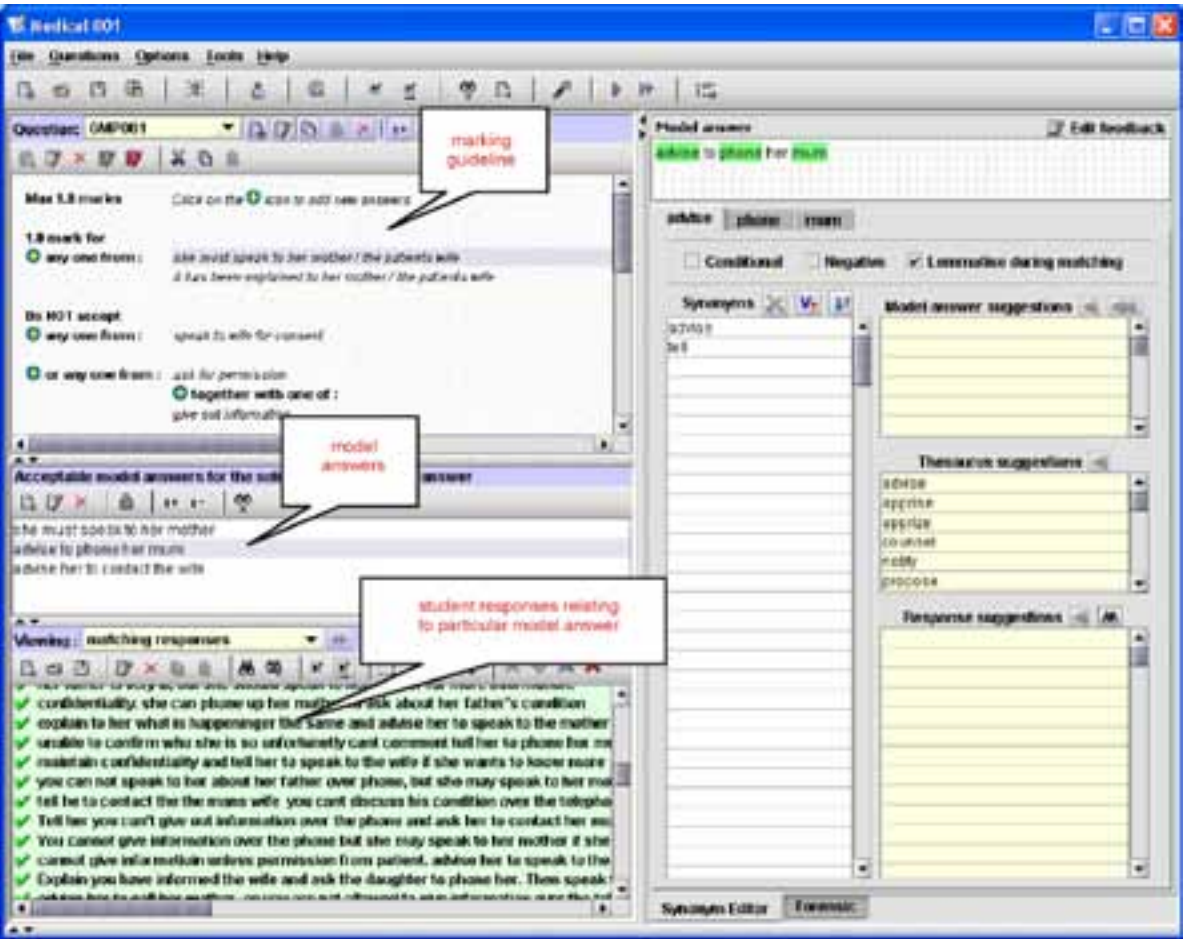
- 3. Mark schemes are uploaded to the server running the marking engine via web services.
- 4. The test is delivered to students (see diagram below).
- 5. Student responses are marked using the computerised mark schemes previously uploaded.
- 6. Optionally, a sample of the student responses may be used to moderate the computerised mark schemes for some items. This is mostly required when new items are being used for the first time. After moderation, responses may be re-marked using the moderated mark schemes
- 7. Results are output.



Human markers mark sample responses against the mark scheme to provide training data for the system during mark scheme authoring.



Authoring the mark scheme in the authoring tool involves an iterative process of adding and modifying model answers, testing the marking at each iteration.



The main mark scheme authoring interface.

Other marking systems

There are also simplified approaches that use a “bag of words” approach and string matching. These can be effective if the teaching context is well-known and consistent restricting the range of likely answers, if the length of answer allowed is restricted and proficient staff are available. It is quite feasible to dispense with the syntactical analysis and simply use string matching of keywords to identify correct and incorrect answers if these rather narrow conditions are met. More sophisticated techniques (though still not employing computational linguistics) such as the OU’s PMatch can be used where series of rules are developed by inspection to act as a correct-answer sieve.

Where next

It is tempting to assume that the marking engines will get more and more sophisticated until they can mark paragraphs of text for meaning but this is unlikely in the foreseeable future. The computer systems do not ‘understand’ the content; they merely check whether the response is a paraphrase of one of the templates linked to the model answer. When human markers mark for meaning, they are comparing what the student says to an internalised world view. As Tom Mitchell says, ‘Computers have no world view.’ The marking engines are ingenious but they are not smart. As the amount of text to be marked increases so does the range of possible responses and the complexity of the marking operation.

Where the skill of weaving content into a sustained argument is required, this will continue to be assessed by human markers, even though achieving high inter-rater reliability is problematic. However, much can be computerised if it is permissible to structure the questions into sub-questions that can be assessed through short text answers..

Where progress will be made is in the improvement to tools for the generation of the (automatic) mark schemes and the facility with which answers can be clustered for human intervention ie increased user-friendliness. While the use of short text marking engines still demands appreciable input and enthusiasm from the human assessor, implying up-front resource costs, the pay off can be considerable – a bank of moderated questions that mark themselves. Indeed, the effort required to create and moderate a mark scheme may be less, in one cycle, than that required to mark by hand. Moreover, academic end-users are rewarded by a greater understanding of the strengths and weaknesses of the questions they have set, the strengths and weaknesses of their students and the effectiveness or otherwise of the instruction.

Benefit

Too often e-assessment is perceived as being coincident with multiple choice testing. The general advantages of e-assessment (viz speed of marking, accuracy & consistency, reduction in marking drudgery, practice for student through repetition, reduced costs, better assessment data collection leading to better feedback) can be extended to other, more natural forms of assessment.

Short answer marking engines offer automatic assessment of constructed answers which may be better than selected response or other closed items at exploring candidates’ knowledge and understanding and thus encourage deeper learning. With more effort, answer-specific feedback can be provided in real time (see OU below).

If a teacher uses questions just once for a single class, there is no point in spending time generating an automatic marking model. But if the questions are used for several classes over several terms the initial effort may be repaid. As the usage increases so does the time saving. The discipline of generating questions that the marking engine unambiguously marks may be valuable for the teacher/assessor.

Furthermore, moderation provides useful insight for the teacher/ assessor into how students actually answer questions – e.g. what their misunderstandings/miscomprehensions are and thereby feedback into teaching and the curriculum. This is a useful side effect of being able to see all student responses for each question listed on a screen.

Open University (OU): short answer free text questions in online interactive assessment

contact details

Sally Jordan **S.E.Jordan@open.ac.uk**
Philip Butcher **P.G.Butcher@open.ac.uk**

The Centre for Open Learning of Mathematics, Computing, Science and Technology (COLMSCT)
The Open University
Walton Hall
Milton Keynes
MK7 6AA
<http://www.open.ac.uk/colmsct/>

Brief details

A small number of short answer questions are included in progress tests which are used both summatively and formatively. There are nine tests throughout the year contributing 20% of overall marks in a 60 credit 1st year undergraduate distance-learning science course (S104). Other items are either selected response (eg multi-choice, drag and drop etc) or type in single word/number answer. Students are given three attempts at each question with an increasing amount of instantaneous feedback. This enables them to learn from the feedback by acting upon it immediately.

What was the problem?

The adult students of the UK Open University study at a distance, usually with the support of part-time tutors who, amongst other things, offer comments and grading on tutor-marked assignments (TMAs). TMAs have always been regarded as having an important teaching as well as assessment function. E-assessment provides the opportunity to deliver feedback to students instantaneously and to free up tutor time to offer support to students in other ways.

The Open University Science Faculty has been using e-assessment for a number of years. The three qualities sought in the feedback on e-assessment tasks are:

- it should be instant
- it should be meaningful, ie sufficiently detailed to allow the student to move on
- it should be contextualised, ie relevant to the mistake made.

The team developing S104 Exploring Science wanted to use regular interactive computer-marked assignments (iCMAs) in addition to TMAs in order to develop understanding and increase student motivation. S104 is a nine-month course that is presented twice each year, with around 1600 students per presentation.

A 20% summative weighting (continuous assessment) was agreed with the university as a compromise between the benefit of encouraging students to engage with the tests and the perceived danger of plagiarism (because correct answers are provided as part of the formative feedback).

The S104 course team wanted to test recall as well as recognition (of the right answer) and also wanted to test students’ understanding and reasoning ability. A spread of question types was used to achieve this.

S104 is a science course containing several disciplines. It was thought that, whilst physics might be adequately assessed with closed questions demanding one word or numerical answers, other disciplines eg biology and Earth sciences, would require more discursive answers. In the event, the physicists also took advantage of the opportunity to ask more searching questions demanding a less predictable short text answer.

In order to mark the short answer free-text responses, the Open University chose to integrate a natural language based assessment engine module from Intelligent Assessment Technologies (IAT) within their own OpenMark system. <http://www.open.ac.uk/openmarkerexamples/index.shtml>

How does the solution work?

The project was funded jointly by the Open University VLE Project and the Centre for Open Learning of Mathematics, Science, Computing and Technology (COLMSCT), as one of a number of practitioner-led projects researching innovative uses of e-assessment in the interactive computer-marked assessment initiative.

The tests are authored, delivered and marked using The Open University’s OpenMark CAA system, an open source system which includes the capacity to deal with constructed responses. This uses mainly xml plus a programming language (Java) for more sophisticated question types and operations. When the Open University decided to adopt Moodle (also open source) as the base system for their VLE, OpenMark iCMAs were integrated within Moodle as an alternative to iCMAs written entirely as Moodle quizzes (the Open University is the global maintainer of the Moodle quiz engine).

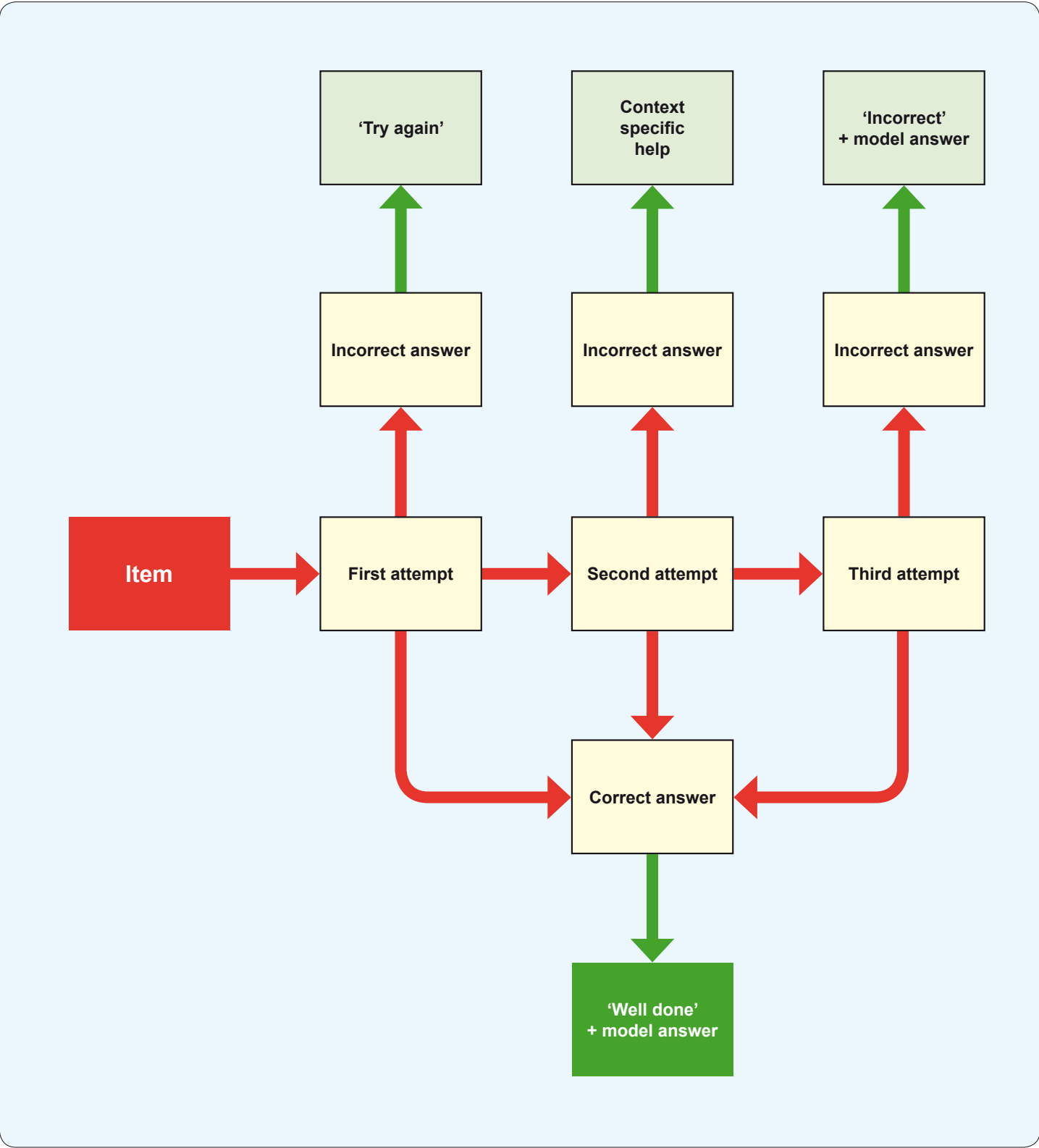
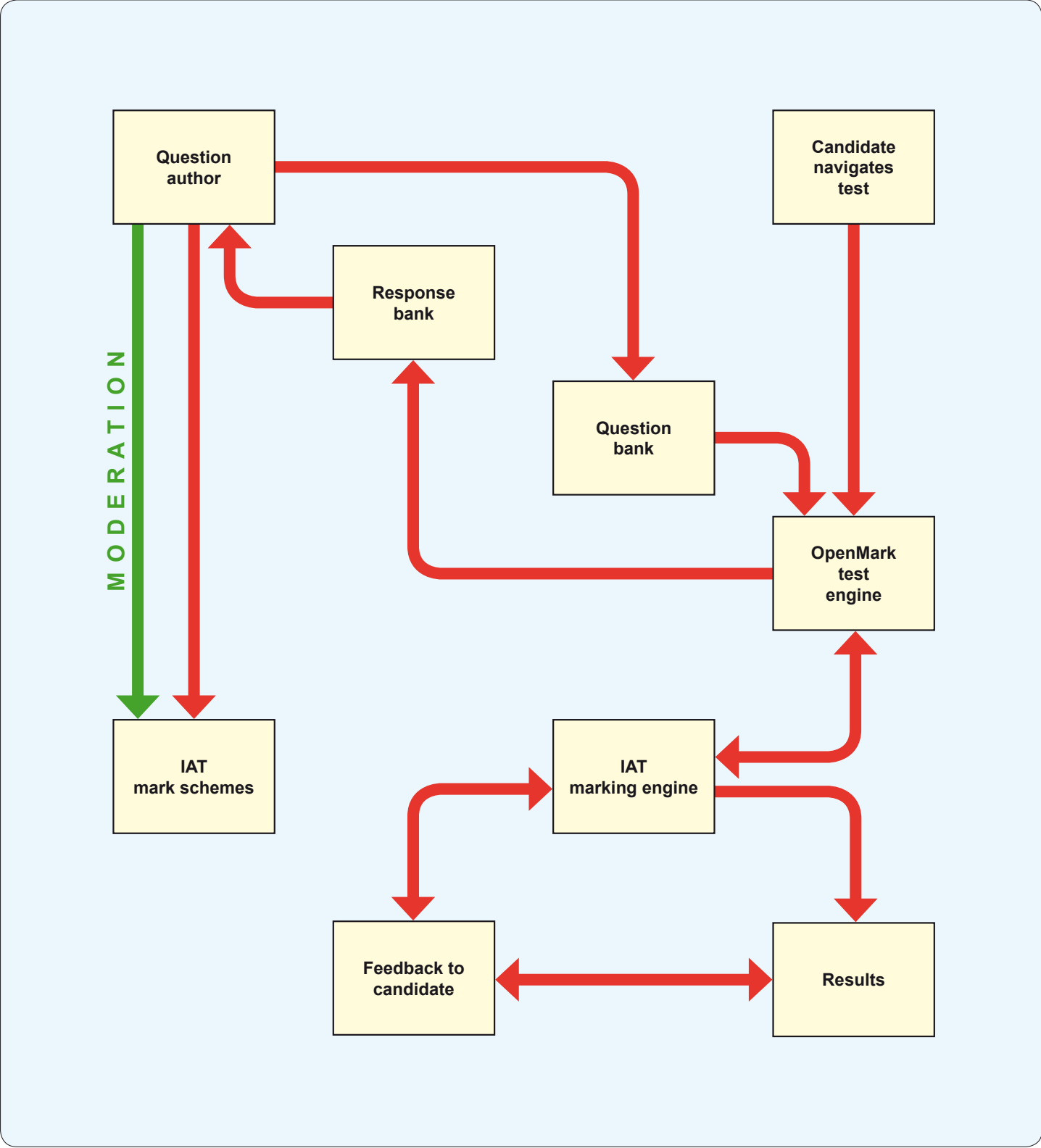
OpenMark provides the university with additional functionality but the appearance and behaviour of tests provided through each system is now quite similar. The IAT add-on allows additional functionality through the marking of short text items

The systems have tended to converge so that it is difficult to tell whether items have been authored using Moodle or in OpenMark (where specialists perform any necessary Java programming)

Because of the problem of plagiarism, many questions are populated with variables ensuring that students get a randomly generated parallel question. This is easy to do in a programming language like java but hard in xml.

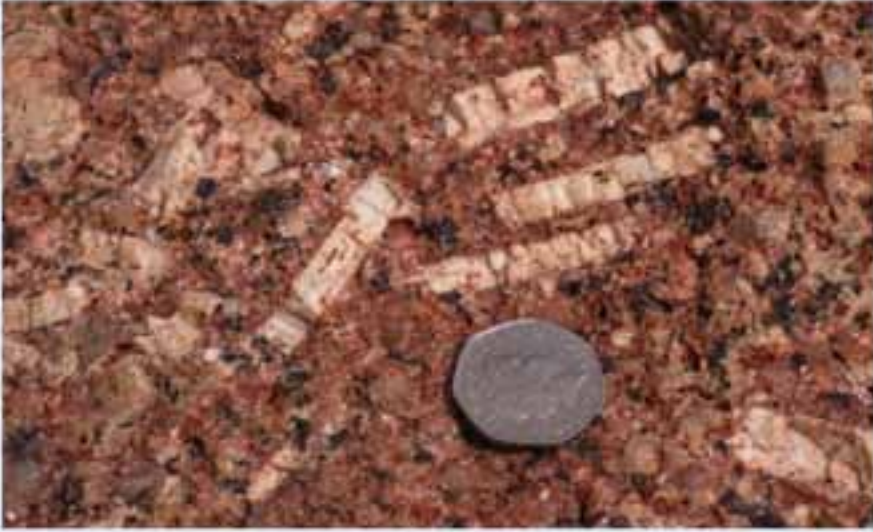
In each S104 iCMA, the student is presented with ten items (and can review them all before attempting any). A correct answer elicits a mark and a model answer. A first incorrect response

usually elicits a general hint (perhaps as simple as “try again”). A second incorrect answer elicits a context specific response, encouraging the student to rethink their approach to the question. A typical iCMA might include around two questions requiring free-text answers of around a sentence in length. The answer matching for these questions is currently written by a member of the course team, using the IAT authoring tool, and each student response is sent from the OpenMark server to an IAT server in order to verify whether it is a correct and to trigger appropriate instantaneous feedback.



Preliminary evaluation indicates that students are highly appreciative of the instantaneous feedback they receive. A student who has recently completed S104 commented, “I thoroughly enjoyed the course especially the interactive DVDs and found the iCMAs a very useful tool for checking that I had understood each topic.”

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 from molten magma

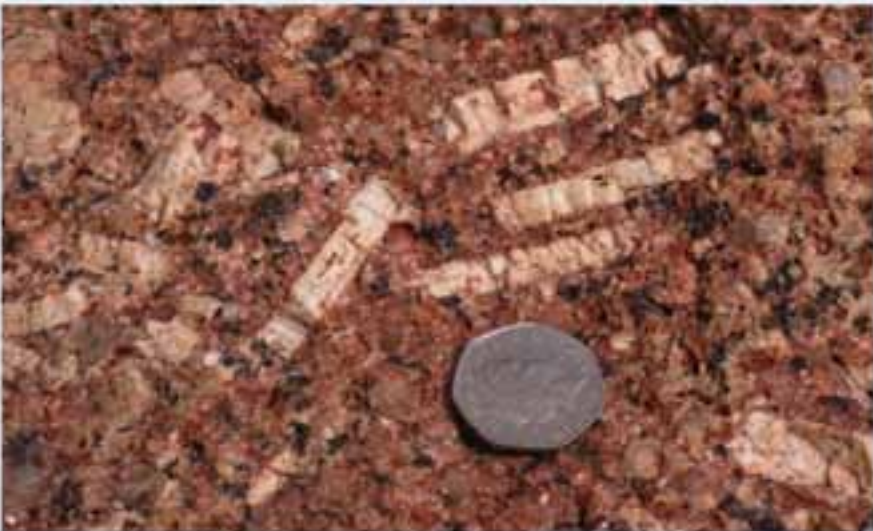
Check

Your answer still does not appear to be correct.

You are on the right lines. Igneous rocks are indeed formed from molten magma. But you need to think about what must happen to the magma in order for an igneous rock to be formed. Then think about what you can deduce from the large crystals visible in the photograph. See Block 3 Section 9.2 and Block 10 Section 5.1.1.

Try again

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 from molten magma which has cooled slowly deep within 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.

Next

Partial credit

Partial credit is awarded when a student gets a question correct after receiving feedback and at their second or third attempt. Student scores are transferred to the Moodle Gradebook and thence to the student and their tutor.

Moderation

The OU prides itself in having brought the system into operation with the maximum number of answers being correctly marked automatically. For open-ended questions, this demands considerable input at the moderation stage when marking schemes are being developed. Reliability in excess of that achieved through normal human marking is routinely achieved.

Benefits

The system works best with questions whose correct answers are fairly predictable. But the OU have purposely been trying to push the system as far as possible and still achieve high reliability. Once trained, the computer marking is generally more consistent than human markers. Questions the computer finds difficult to mark are often ones that human beings also find difficult to mark consistently – eg where the argument is confused or contradictory. This can bring benefits, see below.

The high throughput of students and reuse of questions makes the time spent on generating computer mark schemes worthwhile. In a purely summative system, the reuse of questions might prove problematic given that the system provides correct answers but as their use is predominantly formative, the student is encouraged to work through the test in the way intended and thus generate useful summative scores as well. Clearly this would be impossible in any paper-based system. Responses from students indicate that the feedback appears to motivate students on the course to apply themselves to their learning.

Over time, the system encourages students to express themselves clearly and succinctly, revealing (and honing) their true understanding. Similarly, the moderation process reveals to tutors not only where students have misconceptions but also mistakes in question framing or indeed mistakes in course material.

Future

OpenMark and Moodle iCMAs are being used for diagnostic, formative and summative assessment in an increasing number of situations. Questions using the IAT system are now in use on three courses and members of other course teams have been trained in the use of the IAT authoring tool.

Links and references

For a description of the use of Moodle/OpenMark etc.: Butcher, P. (2008). Online assessment at the Open University using open source software: Moodle, OpenMark and more. 12th International CAA Conference, Loughborough, UK. www.caaconferences.com/pastConferences/2008/proceedings/index.asp

For background on the development of e-assessment at the OU (OpenMark etc): 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.

For the current project:

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

Dundee Medical School

contact details
Professor John McEwen j.mcewen@dundee.ac.uk
Walter Williamson w.m.williamson@dundee.ac.uk
University of Dundee Faculty of Medicine, Dentistry and Nursing
Ninewells Hospital
Dundee
Scotland
DD1 9SY

Brief details

Dundee have produced a fully computerised 270 item progress test, assessing students’ basic core knowledge (recall), the essential knowledge required to be a Pre Registration House Officer (PRHO). All items are short answer and marked automatically. All students in years 1-5 sit the same test in any academic year. On average, each year there is a 20% replacement of items.

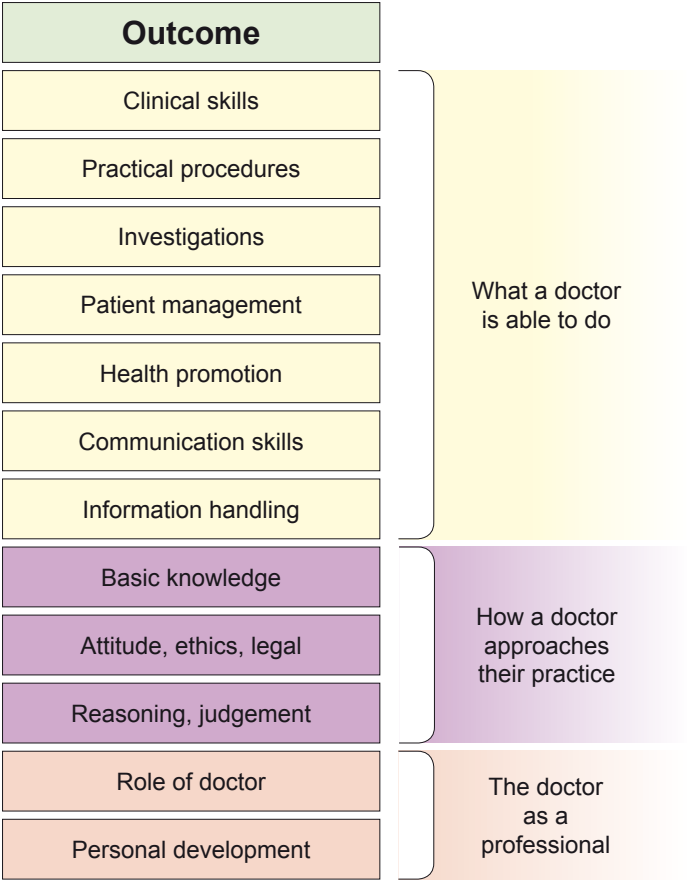
What was the problem?

It was suggested to Dundee by the General Medical Council (GMC) review team that an additional test of essential knowledge be added to the range of assessments employed by Dundee to check the achievements of student outcomes, to provide student feedback and monitor the effectiveness of the teaching. This would provide teachers and students with a sense of how they were progressing relative to their previous performance and to their peers. The use of selected response (ie MCQs) was rejected on the grounds that future doctors would need to determine not select courses of action and that therefore answers should be recalled not recognised. A progress test was designed by the team with the assistance of Professor M. Friedman on secondment from a US university and was piloted on paper in 2001 and 2002.
<http://www.ncbi.nlm.nih.gov/pubmed/10353288>

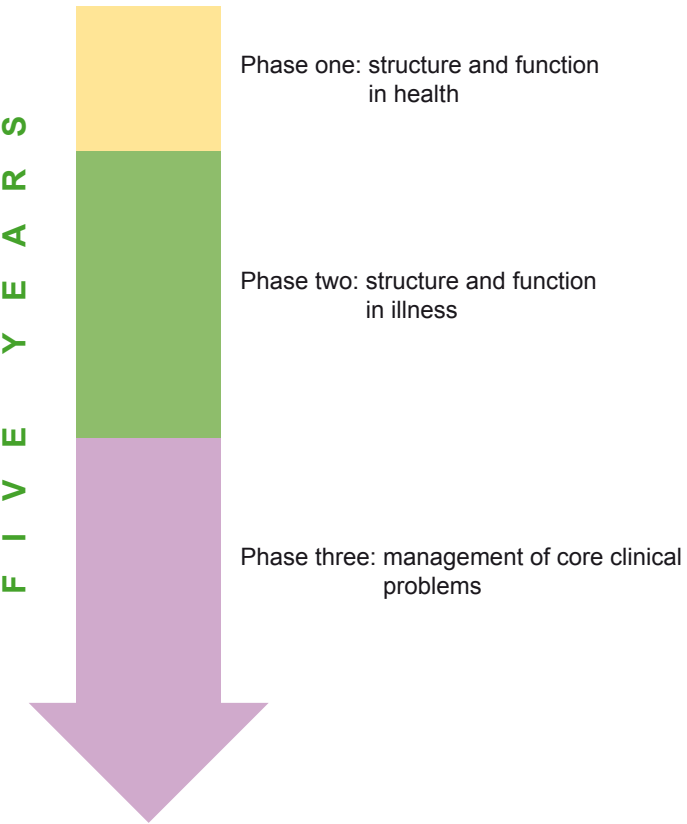
The test was well-liked, reliable and valid but there was a prohibitive workload of marking and moderating in order to provide feedback to the students in years 1-4 and to inform the final assessment in year 5. The decision was then taken to computerise the test.

how does the solution work?

Questions are distributed amongst subject areas in proportion to the amount of curriculum time devoted to them, tagged according to which of the twelve outcomes they assess.



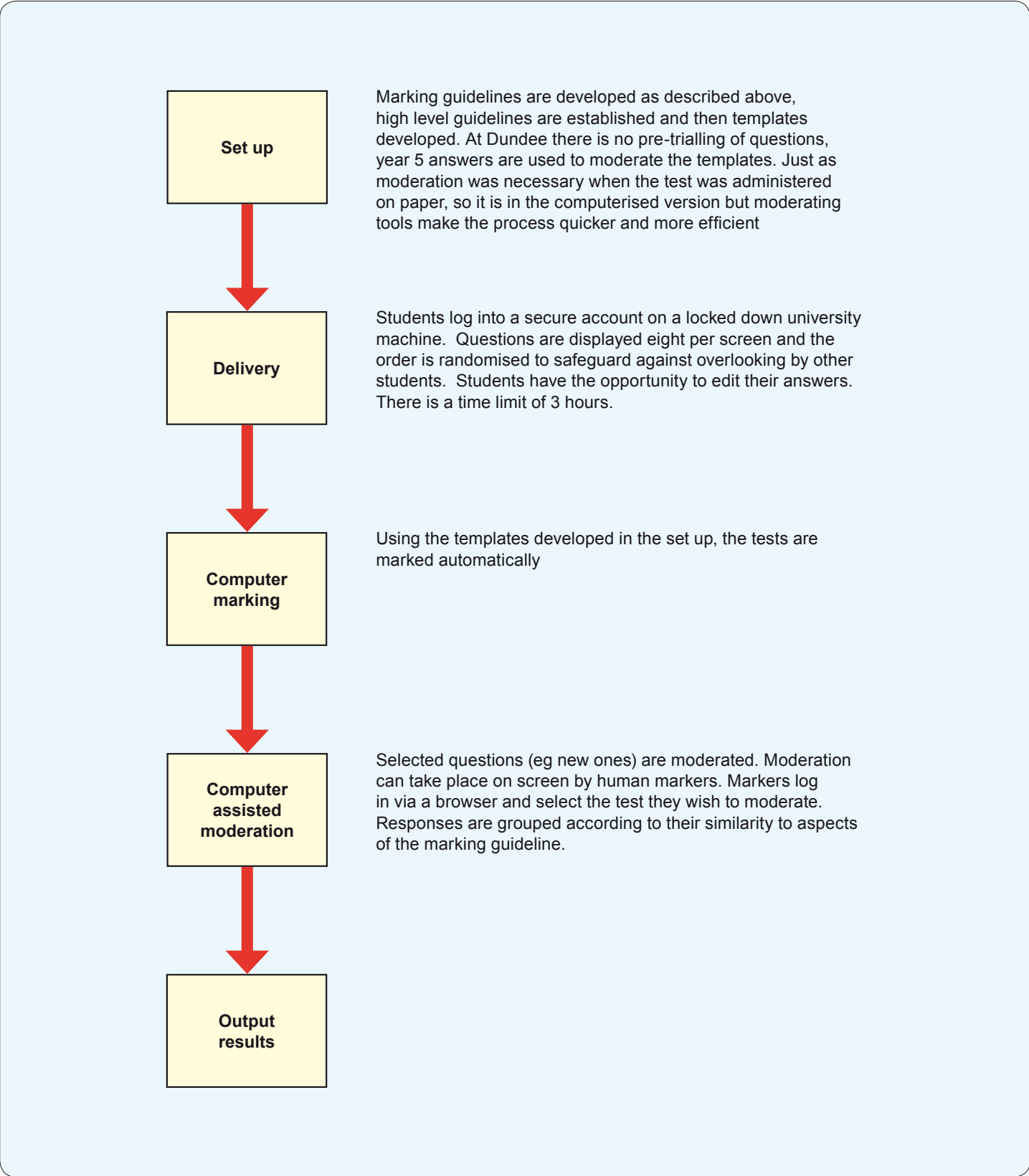
These twelve outcomes permeate all five years of the course. In addition, there are three phases of instruction progressing from the theoretical to the practical.

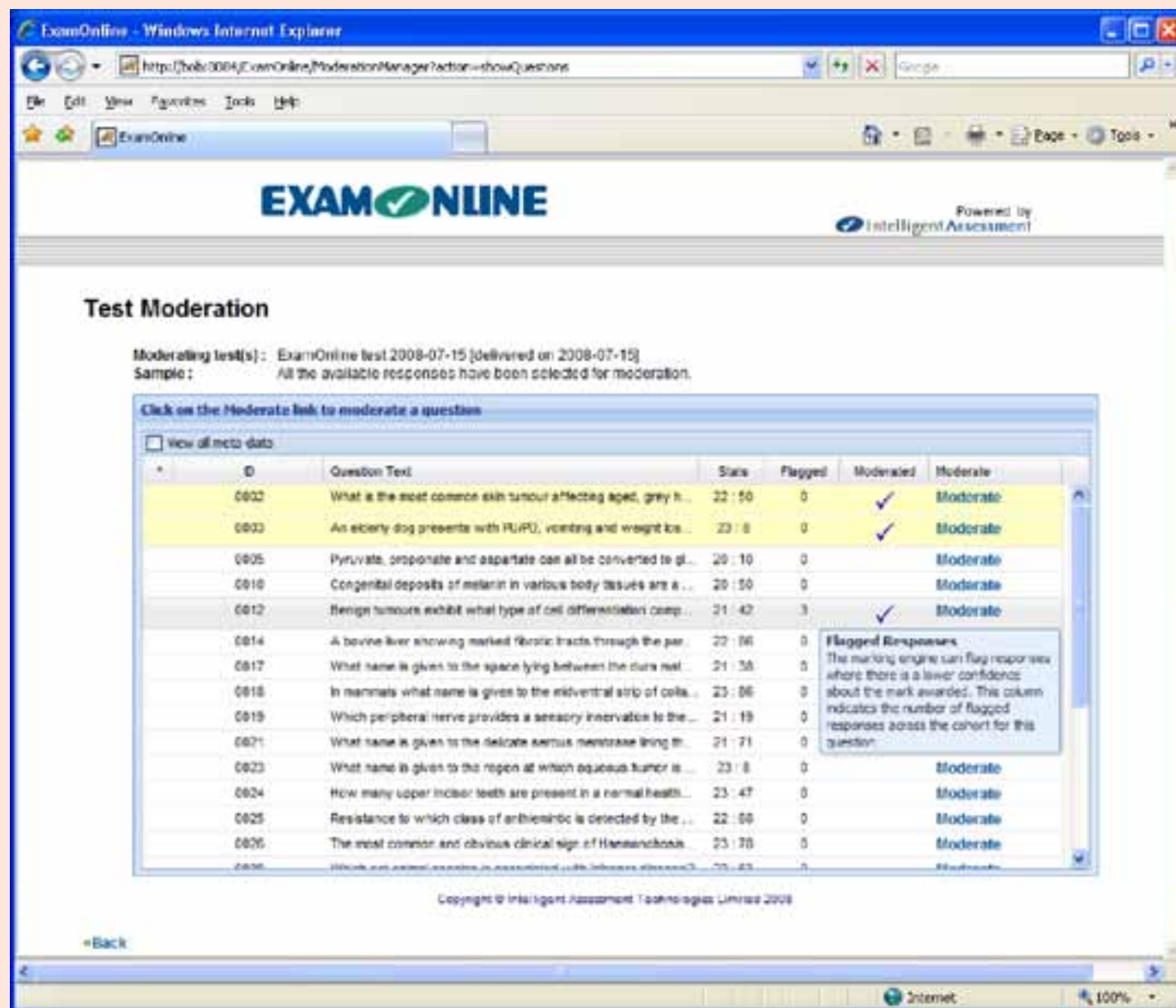


The progress test contains questions dealing with all aspects of the course. Thus first year students have to anticipate the correct answers to much of the material and graduating students need to recall knowledge that was formally introduced years before. As all questions are free text entry, there is far less problem with correcting for guessing. At the end of Year Five, students are assessed on the basis of a portfolio of evidence but, before computerisation, it proved impossible to mark the

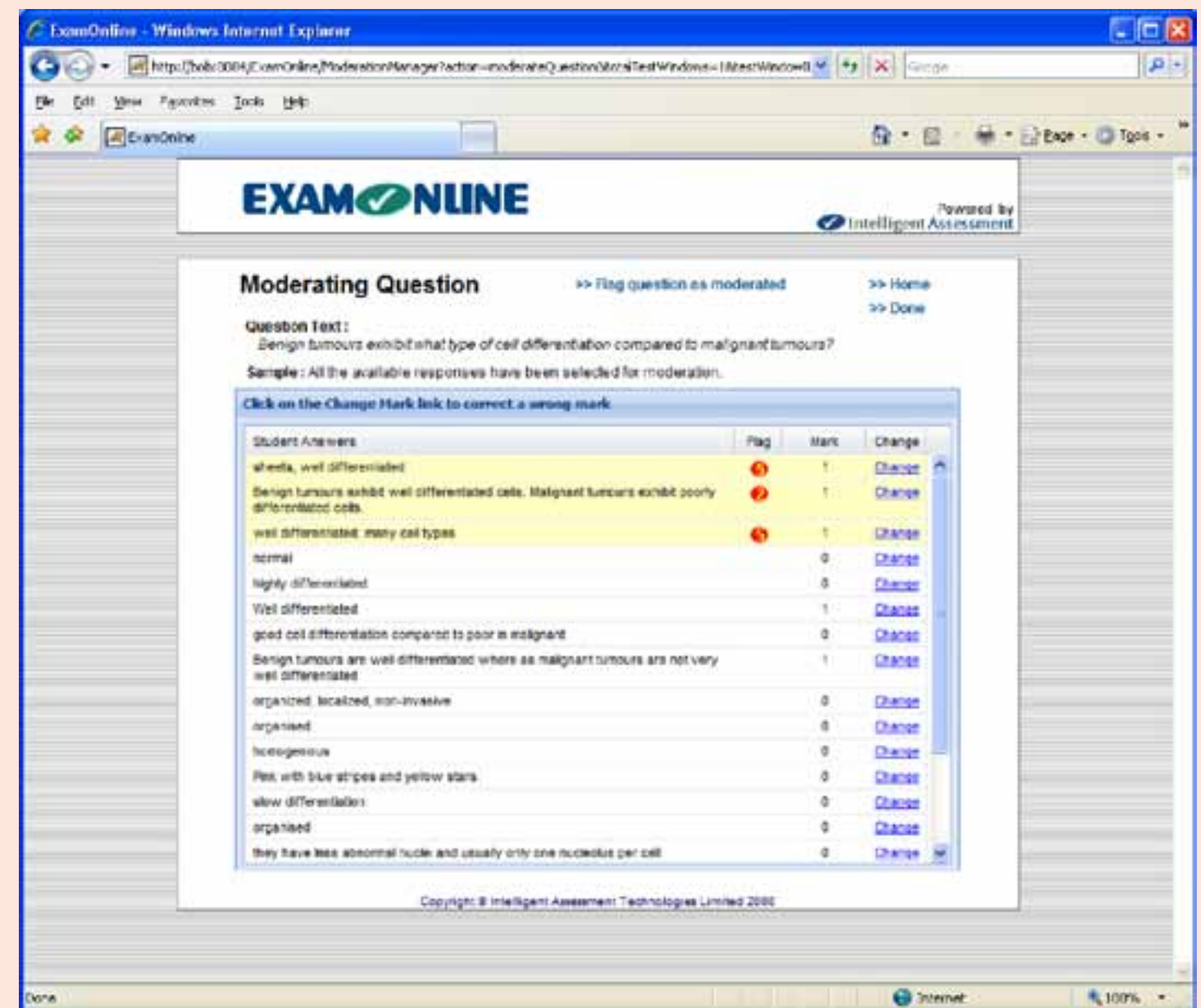
progress test in time for the results to be included. There is no formal pass mark; students are expected to demonstrate mastery of the content of the course.

The answers of the Year Five cohort are used to moderate the marking scheme which is then applied to the later assessments of other year groups. In the early years, paper-based responses were available to the questions.





Selecting which question to moderate. The system provides simple item statistics, and marking confidence indicators from the free-text marking engine.



Moderating a question. The top three responses were flagged by the computerised marking process as possible marking errors (indicated by the red icon in the 'Flag' column).

The experience of computer moderation

Moderation is performed by a group of experts who review the answers to each item together rather than separately, as is necessary on paper. Some corrections are made to computer marking but generally the process reveals problems with the marking guidelines or the item itself. Feedback is available to item writers to help them improve future items.

Accuracy

During the delivery of the initial computerised test, 5.8% of computer marks had to be changed at moderation. However, the majority of these were due to shortcomings in the original paper guidelines (which would also have to have been corrected in a human marking regime). Only 1.6% were due to errors on the computer marking template. After these faults were corrected , the system delivered a match between computer marking and moderated human marking of 99.4%. An analysis of human marking at Dundee revealed between 5 and 5.5% marking errors compared to 5.8% for unmoderated computer marking and around 1% for moderated computer marking.

Developments

Much of the generation of marking templates and corrections to those templates in the light of moderation has been performed by IAT computer staff ie outside the Medical school. From 2009, the new authoring tool provided by IAT will obviate this, allowing the medical school to update its item bank without external support.

Benefits

Computerising delivery has avoided the printing of 800 copies (in multiple versions) of the 30 page test, items are merely uploaded onto the test database. Moderation is quicker and more efficient on screen as responses can be grouped and prioritised. Marking templates can be moderated using a subset of the cohort. Amended schemes can then be applied to the rest of the cohort automatically. This dramatically saves work and speeds up the marking procedures – results can now be generated in time to feed to the end of course assessment process. Years 1–4 can be marked automatically without the need for further moderation, scored immediately and the results fed back to the students.

Students now have access to reliable assessments of their mastery of the twelve outcome areas and can check that they are progressing as they should. Similarly, staff can easily review students' retention and the effectiveness of teaching. In 2005, the teaching model was changed and analysis of the test results, tagged as they are against different outcomes and year of instruction, allowed the effects of the change to be monitored. Transfer students are now routinely given the progress test to determine their level of knowledge.



Short answer marking engines