Computerised Marking of Short-Answer Free-Text Responses.

Tom Mitchell[1], Nicola Aldridge[1] , Peter Broomhead[2]

1. Intelligent Assessment Technologies Ltd. www.IntelligentAssessment.com

2. Dept of Systems Engineering, Brunel University.

Abstract

Open-ended items requiring a free-text response are highly valued in traditional paper-based assessment and learning, but have been absent from computer-based assessment due to limitations in computerised marking technology. Recent developments, however, have seen the introduction of natural language based assessment engines.

One such engine has been developed in the UK by Intelligent Assessment Technologies. The engine looks for specific content within free-text responses, the content being specified in the form of a number of mark scheme templates. Each template represents one form of a valid (or a specifically invalid) answer. The representation of the templates is such that they can be robustly mapped to multiple variations in the input text. The engine has been developed specifically to provide robust computerised marking of short-answer free-text items.

This paper describes the operation of the marking engine, and describes a model based on computerised marking and computer-assisted moderation. A case study for the technology is described, namely the computerisation of a medical progress test at Dundee University, where a test comprising 270 short-answer free-text items is now delivered, marked, and moderated using a computerised system.

Key words : Computer Assisted Assessment, Free-Text, Computerised Marking.

Introduction.

There is now a body of active R&D in the field of CAA of free-text responses, with perhaps the most well-known system being *e-rater* (Burstein, Leacock, Swartz, 2001). In the UK, a project funded by UCLES at Oxford University is aimed at automatically marking GCSE Biology short answers (Pulman, Sukkarieh, 2003).

The system in this paper is based on Intelligent Assessment Technologies' commercially available assessment engine. An early version of this engine was described previously in (Mitchell et al 2002). The engine employs the techniques of Information Extraction (Cowie, Lehnert, 1996) to provide computerised marking of short-answer free-text responses. The system incorporates a number of processing modules specifically aimed at providing robust marking in the face of errors in spelling, typing, syntax, and semantics.

The engine looks for specific content within free-text responses, the content being specified in the form of a number of mark scheme templates. Each template represents one form of a valid (or a specifically invalid) answer.

During the marking process, student responses are first parsed, and then intelligently matched against each mark scheme template, and a mark for each response is computed. The representation of the templates is such that they can be robustly mapped to multiple variations in the input text.

The engine has been employed in projects for the Qualifications and Curriculum Authority (QCA), The Scottish Qualifications Authority (SQA), and Granada Learning.

A Note on Nomenclature.

In this paper the term **marking guidelines** will be used to refer to the (paper-based) keys defined by the item writers which specify acceptable and unacceptable answers for each item. The free-text marking engine used in this project must be configured with a digital version of these marking guidelines (Mitchell et al 2002). In this paper, these are referred to as **computerised mark schemes**.

The term, **moderation** is used to refer to the process of improving and fine-tuning the marking guidelines and computerised mark schemes in the light of 'real' responses given by students. A **moderated mark scheme** (whether paper-based or computerised) is, therefore, one which has undergone the process of moderation.

Marking Short-Answer Free-Text Items by Computer.

IAT's assessment engine employs NLP techniques to perform an intelligent search of free-text responses for predefined computerised mark scheme answers. This is analogous to the process carried out by human markers when marking free-text responses. And like human markers, the system attempts to identify the understanding expressed in a free-text response, without unduly penalising the student for errors in spelling, grammar, or semantics.

The system employs a computerised mark scheme that specifies acceptable and unacceptable answers for each question. The system represents mark scheme answers as syntactic-semantic templates. Each template specifies one particular form of acceptable or unacceptable answer. For example, **Figure 1** illustrates a simple template for the mark scheme answer **The Earth rotates around the Sun**.

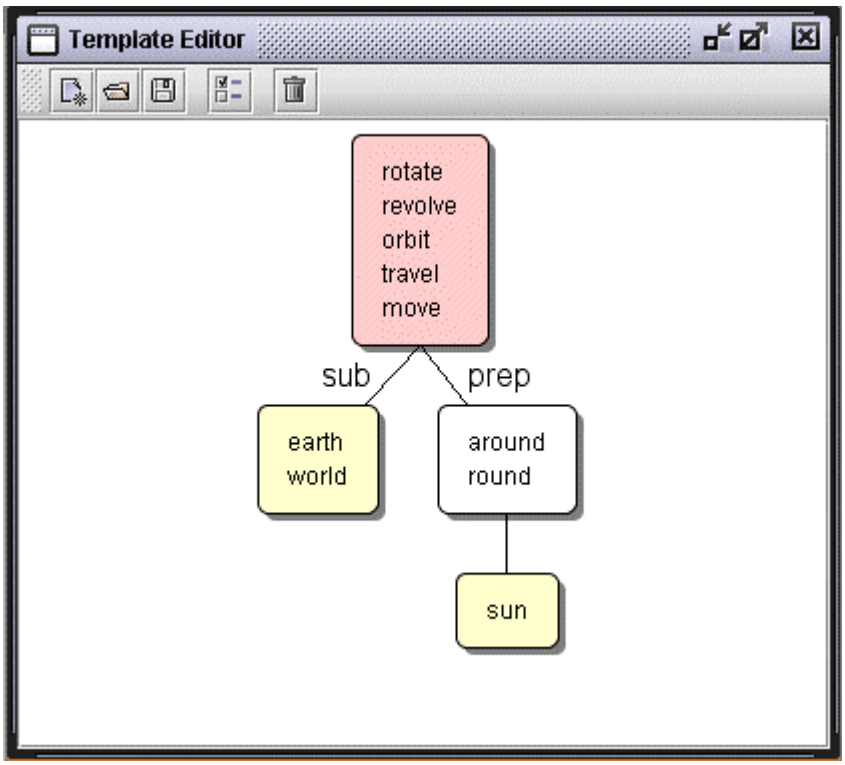

Figure 1. A simple mark scheme template is created using the authoring tool.

The template shown can be expected to match a student response if the response contains one of the stated verbs (**rotate, revolve, orbit, travel, move**) with one of the stated nouns (**earth, world**) as it's subject, and **around / round** the **Sun** in it's preposition.

Verbs in the student response are lemmatised (reduced to their base form, i.e. 'went' lemmatises to 'go') so that, for example, the following student responses will all be matched by the template shown above.

The world rotates round the sun.
The earth is orbiting around the sun.
The earth travels in space around the sun.

Development of the templates in the computerised mark scheme is an offline process, achieved using the authoring tool. The inputs to the process are the marking guidelines for the items, and if available a sample of student responses. The outputs from the process are computerised mark scheme files, as shown in **Figure 2**. The computerised mark scheme files are applied by the marking engine during the marking process, as shown in **Figure 3**.

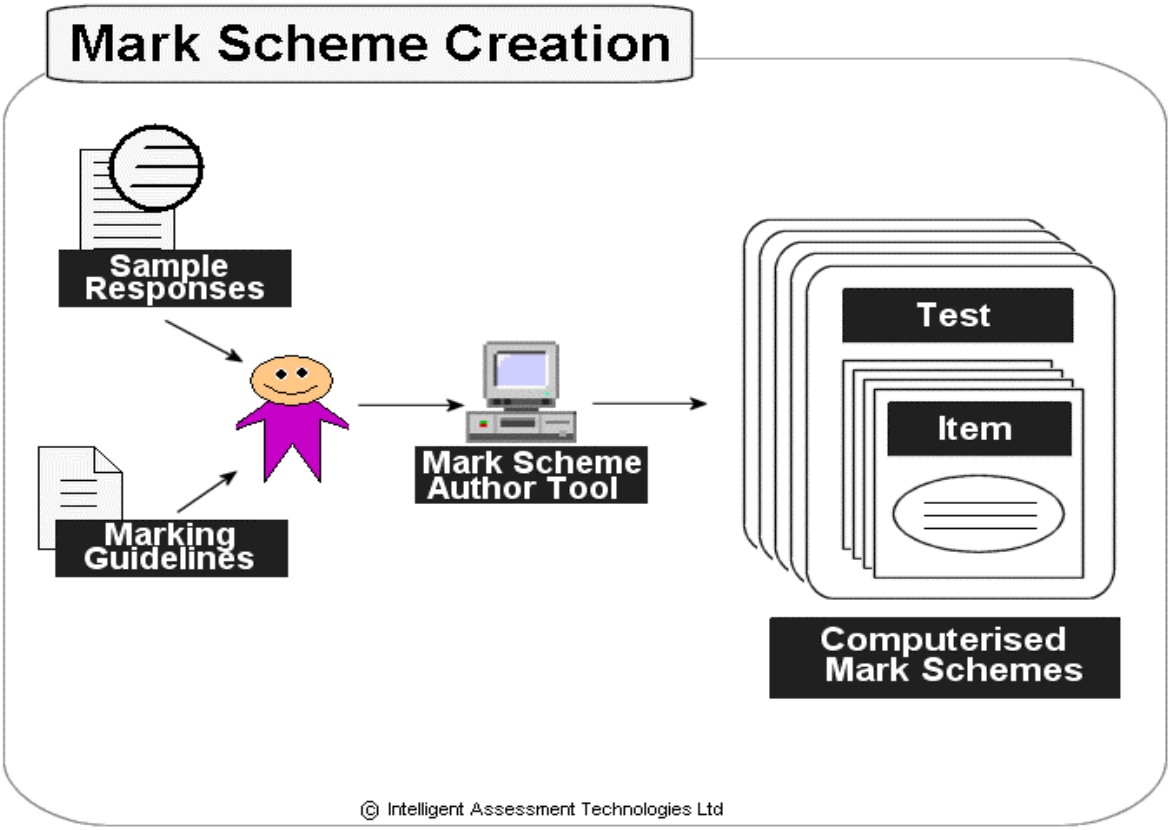

Figure 2. Creating computerised mark schemes.

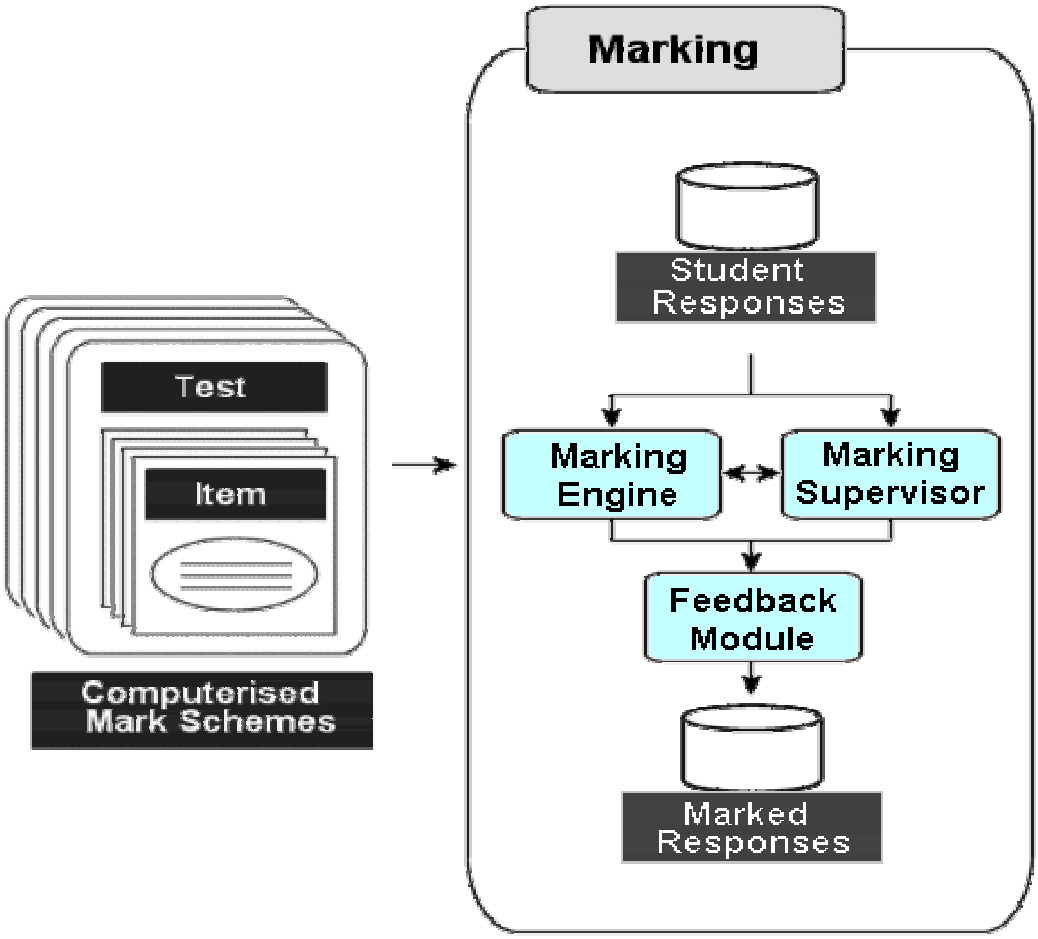

Figure 3. The pre-configured computerised mark schemes are applied to student responses to determine the mark for each response.

Computer-Assisted Moderation.

In conjunction with computerised marking of free-text, we have adopted a model of computer-assisted moderation. In simple terms, computer assisted moderation is implemented as a software interface which enables examiners to efficiently view, and where necessary modify, the marks awarded to individual student responses by the computerised marking process. In addition to this basic functionality however, the interface is developing to provide examiners with sufficient information to enable them to target their expertise where it will prove most valuable, i.e. looking at specific items or individual responses where the computerised marking is likely to be least accurate. The overall approach is to use the marking engine to automatically mark student responses, and then intelligently support the examiners in efficiently moderating the process.

The model has a number of attractive features.

- It allows professional judgement to continue to play a key role in the marking process.

- It changes the emphasis on the way the marking technology is perceived, casting it as a tool which supports the examiners in their job (and increases the efficiency and reliability of the process along the way).

- It allows examining bodies to employ free-text items in computer-based assessments, but to mark them, and moderate the marking, in a more efficient and cost-effective manner than is currently possible.

It may be envisaged that such a system could be used in a number of different ways, for example:

- acting as a ‘second’ marker for comparison with the human marker;
- targeting borderline candidates for re-marking by the human marker;
- highlighting specific responses requiring a human ‘second opinion’;
- improving the skills of item writers and human markers.

A concrete example of the use of computerised marking and computer-assisted moderation is provided in the next section (see Mitchell at al, 2003 for a more detailed exposition).

Computerised Marking and Computer-Assisted Moderation in Practice – The Dundee Progress Test.

The Medical School at the University of Dundee offers a high quality teaching programme, rated Excellent by the SHEFC Quality Assessors. A new assessment, a “progress test” was added to the curriculum in April 2000.

A medical progress test is a comprehensive assessment of medical knowledge designed to inform students about their year-on-year progress against learning outcomes. The test also serves to highlight gaps in their knowledge, and demonstrates their performance relative to their peers. At Dundee the Progress Test is administered annually throughout the five years of the undergraduate programme – each year group sits same test.

Computerising the progress test offered obvious advantages to Dundee, particularly in terms of reducing the marking burden at a time of intense work with summative assessment, and in providing rapid feedback to students. However the progress test itself requires marking of short-answer free-text responses.

During Spring 2003, IAT developed and rolled out a computerised progress test at Dundee. Since the introduction of the system, all five year groups have been successfully tested, a total of approximately 800 students. Computerising the progress test has enabled Dundee to obtain the benefits of CAA, whilst retaining the open-ended item format that they know provides highly reliable and valid tests.

Progress Test Items.

The progress test is comprised of short-answer free-text items. Many of these items can be answered with a single phrase (for example, the name of a treatment or a drug). Others require more of an explanation. Items are written specifically for the progress test, and are not pre-trialled. Some example items are listed in the following table.

Item Text	Marking Guideline
Two days after a myocardial infarction a 50 year old man is found to have persistent fine crepitations (crackles) at both lung bases. What is the most likely cause?	Accept **: Left ventricular failure/ LVF/ Pulmonary (pulm) oedema/ Heart failure/ ventricular failure** Don't accept **: congestive/ right heart failure.**
A 2cm breast cancer (without evidence of metastasis) can be treated by?	Accept **: (Both parts needed): Wide local excision (WLE)/ lumpectomy/ surgery/ lymph node biopsy / excision + radiotherapy/ radiation** Allow**: Mastectomy** Don't accept**: lymph node clearance**
Following haematemesis what basic intervention is required immediately?	Accept**: IV Cannulation/ IV Fluids/ Treat for shock** Allow**: venous cannulation, setting up a drip, giving intravenous fluid therapy or replace/ resuscitation with IV fluids and/or blood or plasma** Allow**: ABC/ Airway, Breathing, Circulation (all 3 needed)** Allow**: venous access** Don't accept alone **: resuscitation/ resuscitate**

The Computerised Progress Test.

The structure of the system developed for Dundee is depicted in **Figure 4**, and described below.

Set-Up.

Set-up is required to configure the free-text marking engine for each item to be marked. Configuration is carried out using the marking guidelines and, if available, a sample of human marked student responses for each item. The output of the set-up process is a computerised mark scheme for each item. These computerised mark schemes are used by the marking engine in the marking process.

Test Delivery.

The first computerised progress test was delivered in April 2003. Sessions were conducted in groups of 80 students at a time in the Universities' IT suite. The sessions were invigilated.

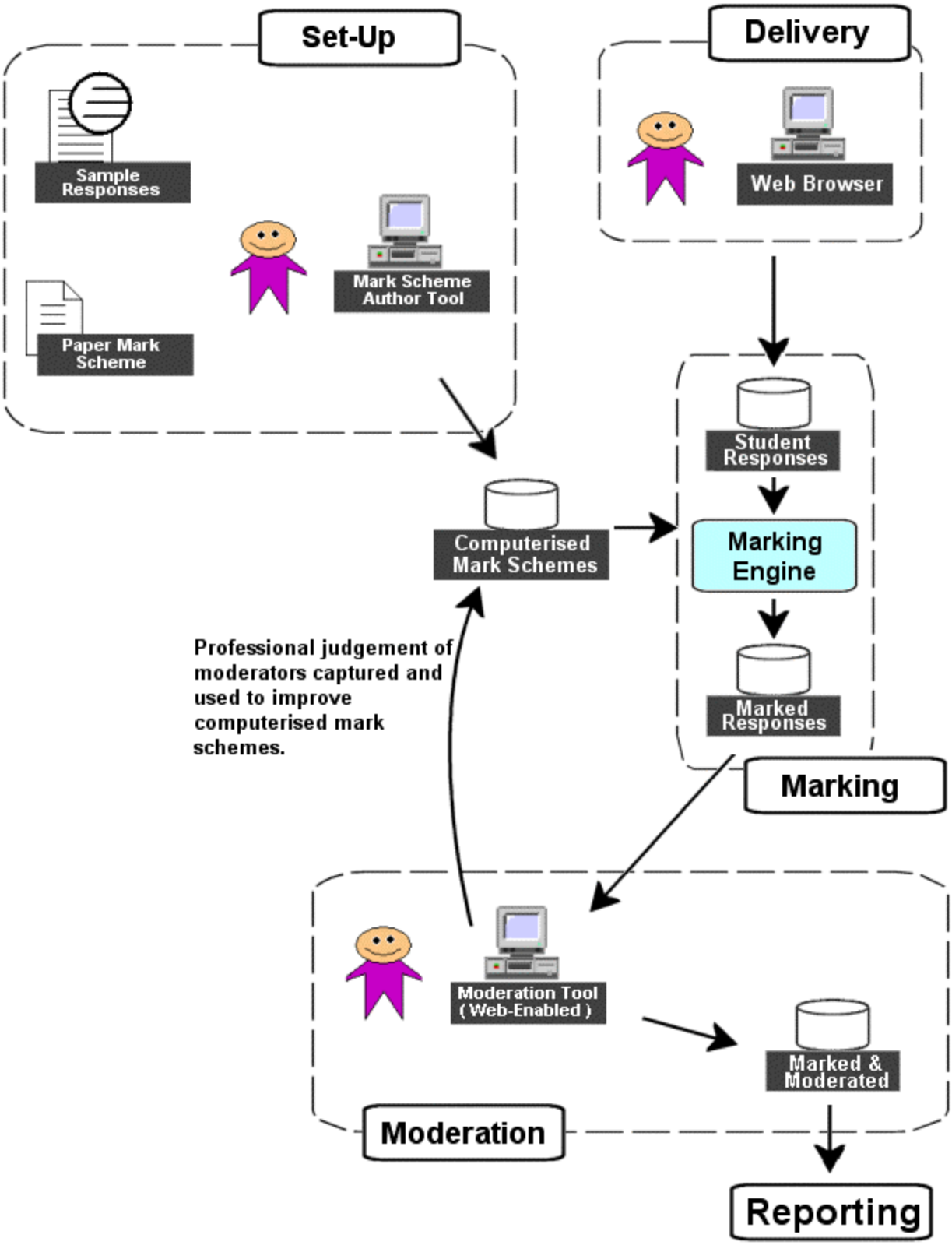

Figure 4. The system developed and rolled-out at Dundee University Medical School in 2003.

The 270 test items comprising the progress test are stored in a database. During test delivery, students are presented with pages of eight items at a time. Item presentation is randomised, such that the probability of adjacent students receiving the same item at the same time is minimised. Upon completion of a page of items, students click on an appropriately labelled button to move onto the next page of items. At this juncture student responses are stored to the database, but are not yet marked. Navigation links are provided to enable students to navigate back and forward through the pages of items, and their responses to previously answered items are displayed when they revisit a page, so that they may, if they wish, edit them for re-submission. Students may quit the test at any time, or the test will end automatically at the end of the 3 hour period.

Between April and July 2003 approximately 800 students sat the computerised progress test.

Computerised Marking.

Marking is carried out in batch after test completion. A simple web interface is provided to enable administrators to select which tests to mark, and to initiate the computerised marking process. The progress of the marking can be viewed, again via a simple web interface.

For the paper-based test, the staff input to support marking for 800 students amounted to approximately **30-man days**. On a 2.4GHz PC running Windows XP, marking takes around 6½ hours for 800 students (approximately 30 seconds per student 'script' of 270 items).

Computer-Assisted Moderation.

The progress test at Dundee is comprised of short-answer free-text items. As with all open-ended items, to obtain accurate and consistent marking the marking guidelines must be moderated in the light of real student responses. The computerised system neither creates nor removes the need for this moderation – computerised mark schemes must be moderated in the same way as paper-based marking guidelines (using a representative sample of the student cohort if pre-trialling has not been carried out). However the computerised system does support and streamline the process.

The items used at Dundee are not pre-trialled, instead the approach adopted at Dundee has always been to moderate the marking guidelines during marking of the Year 5 scripts. The academic calendar at Dundee dictates that Years 2 and 3 are tested in April, followed by Year 5 in May, then Year 1 in early June, and finally Year 4 in early July. With the paper-based test, following on from the Year 5 marking / marking guideline moderation process, the moderated marking guidelines were used to mark the test papers from other year groups. With the adoption of the computerised system the same basic approach is continued – the computerised mark schemes are

moderated using the Year 5 student's responses, and these moderated computerised mark schemes are subsequently used to mark all other year group tests.

The process of computer-assisted moderation is described as follows. Moderators can login via a browser, and select which tests to moderate. They are then presented with a list of all items in the test (see **Figure 5**). Brief item statistics are also presented (the number of students attempting the item, and the percentage awarded a mark). These statistics may be useful in highlighting potential problem items (i.e. where there is an unexpectedly low percentage of students obtaining a mark).

Moderators are able to moderate on an item by item basis. By selecting an item, they can view and change the marks awarded to individual student responses. They can alter the order in which the responses are displayed, so that responses marked as correct / incorrect, and also responses which are similar in length, can be grouped together. This last feature is surprisingly effective at grouping similar answers (see **Figure 6**). Responses for which the moderators change the marks are highlighted in green. Once moderation of responses to an item is complete, moderators can move on to moderating the next item. Previously moderated items are highlighted (see **Figure 5**).

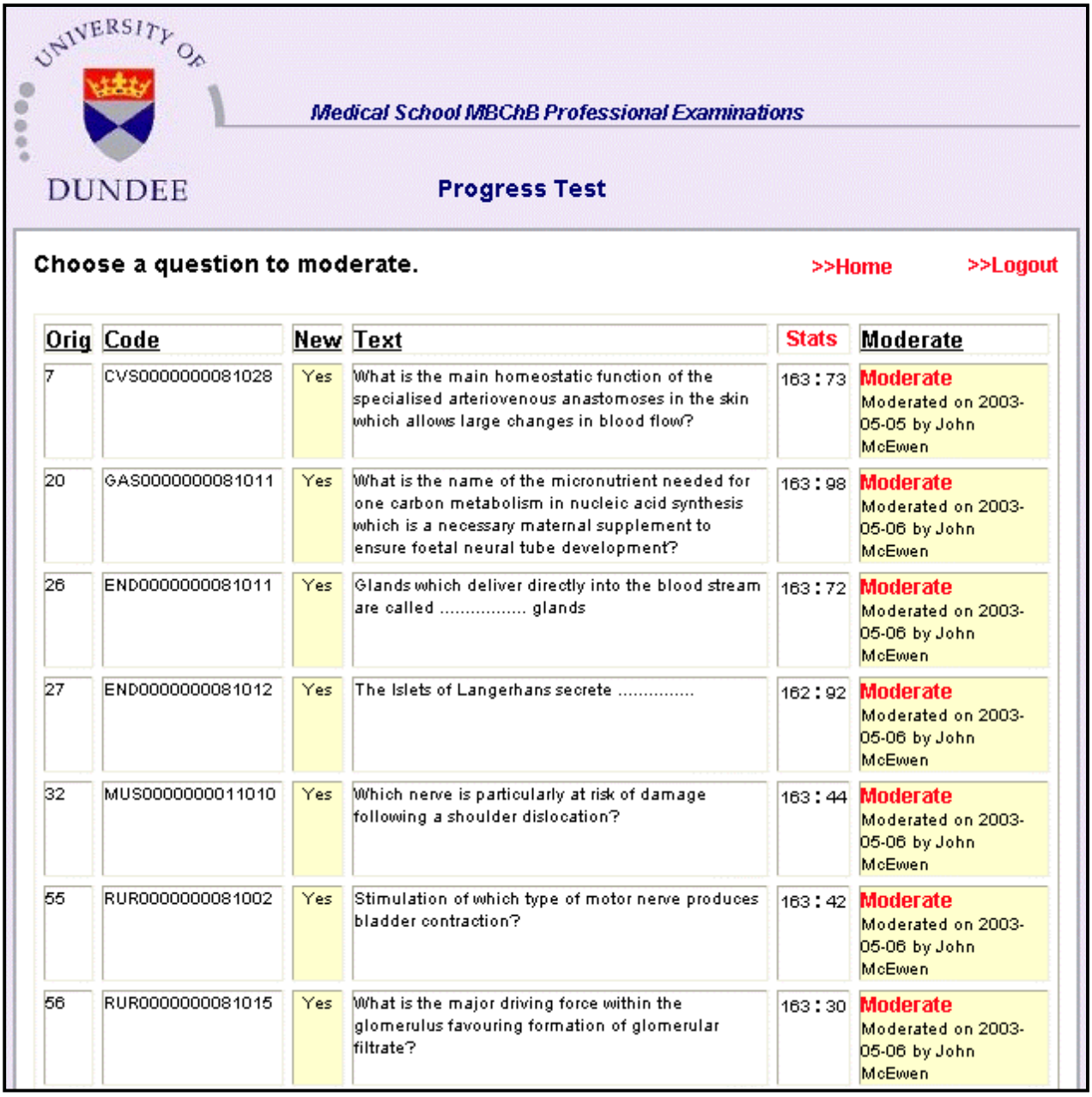

Figure 5. The "menu" page on the moderation interface, allowing an administrator to choose which item to moderate. The "Stats" column gives an indication of student performance on each item, showing the number of students attempting the item, and the percentage awarded a mark.

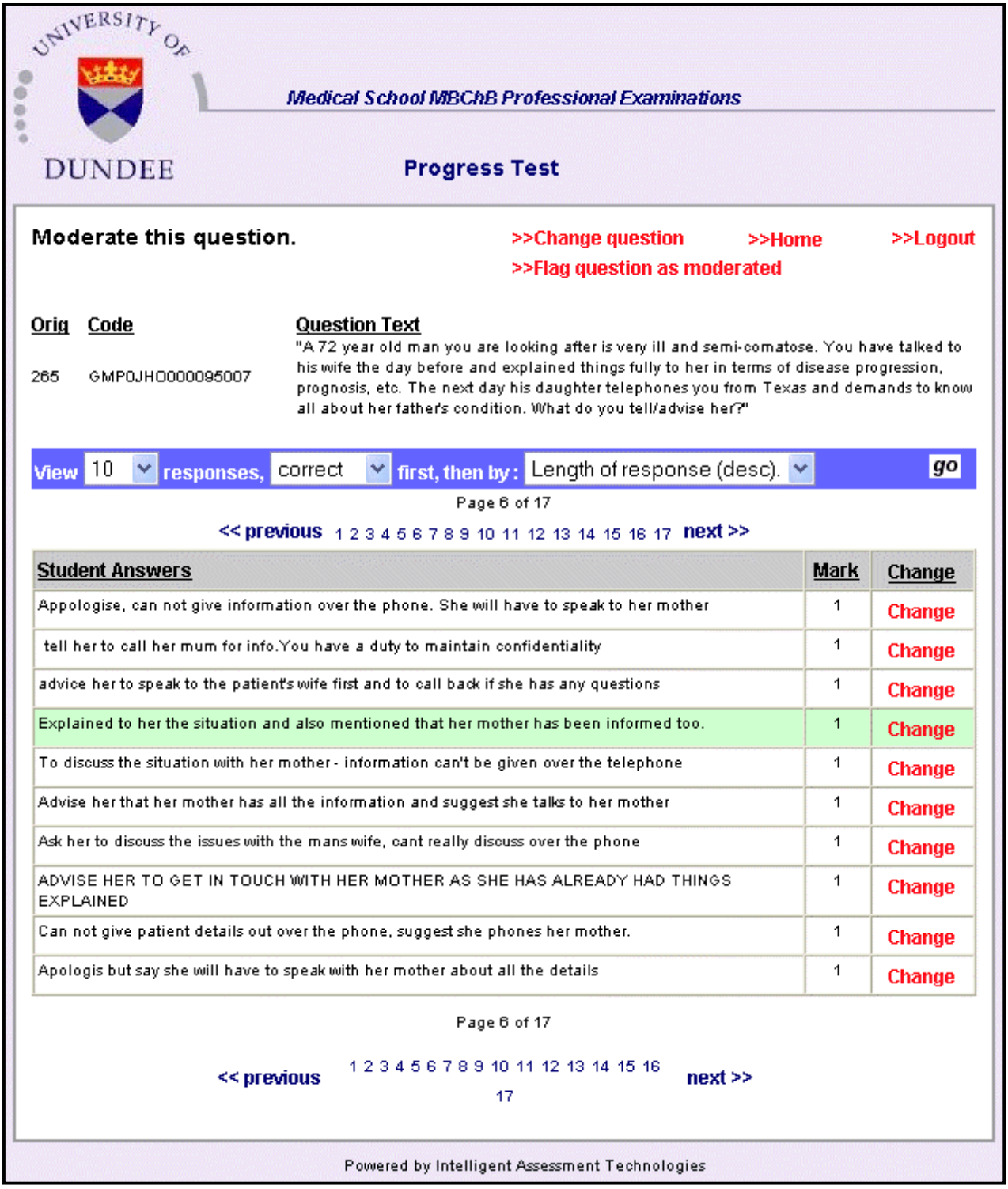

Figure 6. Moderating an item. The marks in the mark column were awarded by computerised marking, and may be amended by clicking on the “Change” link next to the relevant response.

The Accuracy of Computerised Marking in the Progress Test.

Subsequent to the Year 5 moderation process, the computerised mark schemes were re-worked taking into account the changes to the item marking guidelines agreed by the moderation group. The Year 5 test was subsequently re-marked using these moderated computerised mark schemes. The agreement between the computerised marking and the moderated marks resulting from the moderation process was ***99.4%***[1]. The ***0.6%*** error rate is due to system errors inherent in the current version of the marking engine.

Looking at the error rates for individual items reveals that only ***5*** of the ***270*** items had an error rate of ***4%*** or greater – the worst being ***7%***. For each of these "problem" items, the marking guidelines are quite broad and unspecific. Such items are typically difficult to mark consistently, either by computer or human. With the computerised system however, such items can be efficiently targeted for human attention.

Validating the Computerised System.

The moderated computerised mark schemes were thereafter used to mark all other year group tests (these had not been part of the moderation process). To check the accuracy of the marking, 10 Year 2 and Year 3 students were selected at random[2]. Their responses were hand marked using the moderated marking guidelines, and the results compared with the marks awarded by computer using the moderated computerised mark schemes. The results of this exercise are summarised below.

Number of Students Affected	**Marks Gained / Lost by Hand Marking**
5	0
4	+1
1	+2

As can be seen from this table, the computerised marking errors tend to be missed positives rather than false alarms. One mark difference in a student's score equates to an error in the student's percentage of ***0.37%***, two marks to ***0.74%***. From the 10 students selected at random therefore, the mean error in their percentage scores is ***0.22%***, with the highest being ***0.74%***.

As a final check, a small selection of Year 5 students were selected for human versus computer marking. These students were not chosen at random, but rather were picked from students who had done either significantly better or worse in Year 5 than they had in Year 4 (indeed one student requested his mark be checked). The students' responses were printed out, and hand

[1] The accuracy achievable with computerised marking of short-answer free-text varies from item to item. The items in the progress test are generally very suitable for computerised marking, hence the very high marking accuracy.

[2] This may seem a small sample, but in fact represents over 2,600 responses.

marked using the moderated marking guidelines. No discrepancy between the computerised marking and the human marking was encountered.

Computerised Marking versus Human Marking.

The progress test is particularly onerous to hand mark. There are approximately 800 scripts, 270 items per script, and a team of 6 markers can together mark around 15 scripts per hour. Furthermore, the marking guidelines, although detailed and (usually) prescriptive, can be difficult to apply consistently.

In two separate exercises, the error in the hand marking at Dundee for the paper-based tests has been measured at between ***5*** and ***5.5%***. This is comparable to the marking error obtained with unmoderated computerised mark schemes (***5.8%***). With the moderated computerised mark schemes, the marking error is substantially lower (of the order of ***1%***).

For this test at least therefore, system errors inherent in the free-text marking engine (Mitchell et al 2002) are less significant than errors in human marking, where differences in interpreting marking guidelines, inconsistencies in applying agreed marking guidelines, the effects of tiredness, and of course simple human error routinely play a part.

Benefits of the Computerised System.

The main advantages of the computerised system include the following.

- The marking burden on staff was eliminated, and the marking turn-around time was greatly reduced. For previous years' paper-based tests, the staff input to support marking of 800 student scripts amounted to approximately **30-man days**. On a 2.4GHz PC running Windows XP, computerised marking takes around 6½ hours for 800 students (approximately 30 seconds per student 'script' of 270 items).

- The academics felt that the process of moderation via computer was a largely positive experience. Academics can easily detect weaker items, with the additional advantage that collated student responses give insight into curriculum coverage. There was a common view that item-writers should be involved in future moderation meetings, as it would help them produce better items.

- On-screen moderation was quicker than expected. Responses could be scanned quickly, and most items required little input. The moderated items can be re-used in future tests, with a high level of confidence in the computerised marking.

- As always with CAA, flexibility is increased. Already at Dundee, students who were unavoidably absent on the day of a test (due to

illness or work placement) have been able to sit the test with virtually no admin burden for Dundee staff.

Conclusions

This paper has described a marking engine specifically developed to provide robust computerised marking of short-answer free-text items. The engine looks for specific content within free-text responses, the content being specified in the form of a number of mark scheme templates. A model based on computerised marking and computer-assisted moderation has been described, and illustrated by reference to an implementation at Dundee University. The technology described can enable examining bodies to obtain the benefits of CAA, whilst retaining short-answer free-text items which they know provide highly reliable and valid tests.

References.

Burstein, J., Leacock, C., Swartz, R., (2001) Automated Evaluation Of Essays And Short Answers. Fifth International Computer Assisted Assessment Conference Loughborough University 2nd and 3rd July 2001.

Cowie, J., Lehnert, W.G. (1996). Information Extraction. In Communications of the ACM vol. 39 (1), pp. 80-91.

Mitchell, T., Russell, T., Broomhead, P., Aldridge, N. *Towards Robust Computerised Marking of Free-Text Responses*, Proceedings of the 6th International Computer Assisted Assessment Conference, Loughborough 2002, pp233-249.

Mitchell, Aldridge, N., Williamson W., Broomhead, P. *Computer Based Testing of Medical Knowledge*, Proceedings of the 7th International Computer Assisted Assessment Conference, Loughborough 2003, pp249-267.

Pulman, S., Sukkarieh, J. *Automatic Marking of Short Answers.*, presentation at Scottish Centre for Research into Online Learning and Assessment (SCROLLA) Free-Text Analysis Symposium, Heriot-Watt University, April 2003.