Computer Based Testing of Medical Knowledge.

Tom Mitchell[1], Nicola Aldridge[1],Walter Williamson[2], Peter Broomhead[3]

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

2. Faculty of Medicine, Dentistry and Nursing, University of Dundee.

3. Dept of Systems Engineering, Brunel University.

Abstract

The Medical School at the University of Dundee offers a high quality teaching programme, rated Excellent by the SHEFC Quality Assessors. The assessment of an outcome based curriculum, and the need to provide rapid student feedback, represents an ongoing challenge for the School. The recent introduction of a "progress test" has only added to these challenges. Computerising the progress test offers 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 free-text responses. Objective testing is not an acceptable alternative.

This paper details the development and roll-out of a computerised system for delivering and marking the progress test in the Medical School at Dundee. The system employs an innovative natural language based assessment engine. The assessment engine has been developed to perform robust computerised marking of free-text responses to open-ended items.

The progress test consists of 270 short-answer free-text response items. Item presentation is randomised, such that the probability of any two students receiving the items in the same order is negligible. Students are allowed up to three hours to complete the test.

Computerised marking of student responses is carried out in batch mode, once the test is complete. The system provides a simple interface to enable administrators to initiate computerised marking of student responses, and to provide information on the progress of marking. The system supports moderation of the marks awarded by computer. Results of the tests are exported in a flat file format for subsequent processing and reporting.

The paper details the experiences gained in testing over 450 medical students so far at Dundee in 2003. A comparison with the previous years' paper-based testing approach is provided.

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

Introduction.

The Medical School at the University of Dundee offers a high quality teaching programme, rated Excellent by the SHEFC Quality Assessors. The assessment of an outcome based curriculum, and the need to provide rapid student feedback, represents an ongoing challenge for the School. The recent introduction of a "progress test" has only added to these challenges. Computerising the progress test offers 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 free-text responses. Objective testing is not an acceptable alternative.

This paper details the development and roll-out of a computerised system for delivering and marking the progress test in the Medical School at Dundee. The system employs an innovative natural language based assessment engine. The assessment engine has been developed to perform robust computerised marking of free-text responses to open-ended items.

CAA of Free-Text Responses.

There is now a body of active R&D in the field of CAA of free-text responses.

Perhaps the most well-known system is *e-rater* (Burstein, Leacock, Swartz, 2001), an automatic essay scoring system employing a holistic scoring approach. The system is able to correlate human reader scores with automatically extracted linguistic features. More recently ETS have developed *c-rater*. C-rater evaluates the accuracy of written responses in terms of scoring rules that define the requirements of a full or partially correct answer (ETS 2003).

An ambitious approach which appears to show high promise is that of Latent Semantic Analysis (LSA) (Landauer, Dumais, 1997). LSA has been applied to essay grading, and high agreement levels obtained (Landauer, Foltz, Laham, 1998).

In the UK, a project funded by UCLES at Oxford University is aimed at automatically marking GCSE Biology short answers (Pulman, Sukkarieh, 2003). Initial experiments have been carried out on 206 marked student answers using Information Extraction (IE) tools (specifically, a part of speech tagger, and a noun phrase and verb phrase chunker). Promising initial results have been obtained, but the research is now focused on developing an example based classification system.

The system in this paper is based on the commercially available AutoMark engine developed by Intelligent Assessment Technologies and described previously in (Mitchell et al 2002). The AutoMark engine employs the techniques of Information Extraction to provide computerised marking of short 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. AutoMark 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. 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. AutoMark has been employed in projects for the Qualifications and Curriculum Authority (QCA), The Scottish Qualifications Authority (SQA), and Granada Learning, and can now be integrated with QuestionMark Perception.

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 Progress Test.

Background.

The General Medical Council (GMC) produced a policy document on Medical Education called "Tomorrows Doctors" (GMC, 1993). It set guidelines for "core" knowledge, defined as "essential knowledge a student required to practice in the Pre-Registration House Officer (PRHO) year". The success of Dundee's adaptation to those guidelines has been measured by external bodies such as the GMC, with the award of an "Excellent" grade at the quinquennial medical course review.

Dundee established competencies in twelve outcomes necessary to be a good doctor. Although not as directly stated in the GMC document, they mirrored the spirit and content. Assessment of the course was carried out at the end of each year of study with an examination format common to most medical schools involving a written component consisting of an Extended Matching Item (EMI) and Constructive Response Questionnaire (CRQ), and a second clinical practical called an Objective Structured Clinical Examination (OSCE). Against this background the GMC Review Team suggested an additional assessment should be considered to underpin those written components to assist in student feedback, outcome achievement and course audit. This assessment should not necessarily be a summative test, but would provide coverage of all aspects of "core" knowledge and in so doing identify to the student and their tutoring staff the individual students relative position concerning outcome level achieved within a given year, and year by year.

Through the assessment experience and expertise of Professor M. Friedman, on secondment from a US University, a "progress test" was designed to address this perceived weakness in the assessment process. The first pilot was initiated between April 2001 and June 2001 involving each of the five years at appropriate conclusion to their years' studies. The item styles were completion statements, definitions and short-answer. Each item was appropriately coded to identify its content year, body system, curriculum outcome, core clinical problem and clinical block. Although the Multiple Choice Question (MCQ) format is broadly used throughout medical schools (Fowell, Bligh, 1988) to assess competency it was discounted because it was felt that "a doctor does not get presented with five choices" (Veloski et al, 1999) when encountering a patient. Also MCQ's were strongly dismissed because advice was given that it was most important to ensure the exam was one of "recall" not "recognition". The choice of format was also reinforced by the weight of evidence that many schools in the USA were moving to an open-ended response question format.

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. Some example items are listed in the table below.

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**

Items are written specifically for the progress test, and are not pre-trialled.

Problems of Paper-Based Testing.

The Year 5 progress test had, due to the academic calendar, to be scheduled within 10 days of the Year 5 final exam. With the paper-based system, results were not available until after that event. Consequently the objective of providing feedback or assessment to inform the final examination was lost. Due to the intensity of work involved in marking, scoring and reporting this delay prevailed for all cohorts (there are approximately 160 students in each of the five year groups). In total for all examinations the staff input to support those activities amounted to approximately **30-man days**, *exclusive* of academic preparation of items, standardising content and production of the 30 page exam scripts. The pilot test alerted the medical school to a number of issues that was delaying achievement of the initial aims of the progress test. However, more positively the pilot had shown that the progress test was a highly reliable and valid test. In addition the student body welcomed the test recognising that feedback was not yet in a timely manner, but when resolved would be a worthwhile assessment for feedback and structured learning.

The second pilot in 2002 encountered the same problems with no resolution to a paper format and the management required to deliver it. The number of items was increased to 270 to reflect more broadly the final year of study. However the delays remained and were unavoidable where such an input of staff time was required. It was recognised at the review of the Test that the major objectives of the test, firstly to inform the final year students and examiners prior to the Portfolio Examination, and secondly provide timely feedback and progression information to the earlier years were not being achieved. Also pressure on academic staff time, ability to moderate and audit examination output, production costs and pressure on assessment administrative staff at a time of intense work with summative assessment indicated the test in its paper format was unsustainable.

A Pilot.

The potential benefits of computerising the progress test at Dundee were obvious, particularly in removing the very significant marking burden. In autumn 2002 a pilot project was carried out to prove the feasibility of delivering and automatically marking progress test items. The objectives of the pilot were to :

- assess the reaction of the students to a computerised progress test;
- examine the accuracy of computerised marking for progress test items;
- contribute towards defining the specification of a full system.

Twenty five items were selected for the pilot from the previous years' progress test. Intelligent Assessment Technologies were supplied with the paper-based marking guideline for each item, and approximately 50 marked student scripts containing the pilot items. The paper-based marking guidelines and the sample student responses were used to develop the computerised mark schemes required by the marking engine (Mitchell et al 2002). The test was assembled and delivered using web-based technology, so that the students accessed the computerised test using a standard web browser from a University IT suite. In two sessions in late November / early December 2002, approximately 30 students attempted the 30 minute 25 item test. The student responses were subsequently computer marked, and then the computerised marking was manually checked. The error rate for the computerised marking across the 25 items was approximately *1%*. Student feedback from the pilot was very positive. The decision was taken to move onto a full system. However the pilot also identified additional functionality required for the full system – the ability to check the marks awarded by the computer, and if necessary change them. This ability to "moderate" the marks awarded by the computer became a key ingredient of the full system. The issue of moderation is discussed briefly in the next section.

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. This is particularly true for "new" items – i.e. those items which have not been used previously. How this is achieved at Dundee is detailed below.

The Moderation Process.

The items used at Dundee are not pre-trialled, instead the approach adopted at Dundee has been to moderate the marking guidelines during marking of the Year 5 scripts. The academic calendar 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. Following on from the Year 5 marking / marking guideline moderation process, the moderated marking guidelines are used to mark the test papers from other year groups.

Year 5 students are likely to provide a full range of acceptable answers, and so enable proper moderation of the marking guidelines. However this marking / moderation had previously been an onerous process. The 160 or so Year 5 scripts were marked over a number of days by a group of senior academics representing a broad range of expertise across the learning outcomes. Student scripts were marked one at a time, with the panel of academics passing judgement on each of up to 270 responses per script. During the marking process, the decisions on whether or not to accept particular student responses were also used to moderate the marking guidelines – typically this means adding additional acceptable answers to the marking guideline. At the end of marking the Year 5 scripts the marking guidelines are deemed to be fully moderated and complete, and are then used to mark the scripts from other year groups.

The computerised system neither adds to nor removes the need for 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. This is detailed in the later section on moderating the computerised tests.

The Computerised Progress Test.

The structure of the full-system is depicted in Figure 1, 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.

Students log into an "examination account" – essentially a tied-down Novell account which automatically runs a web browser in kiosk mode with the progress test URL. No other programs are available on the desktop from the exam account.

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.

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.

Computer-Assisted Moderation.

The system supports moderation of the marks awarded by computer. 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, with the list sorted such that "new" items (i.e. those items which have not been used in previous years' tests) are at the top (see Figure 2). 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).
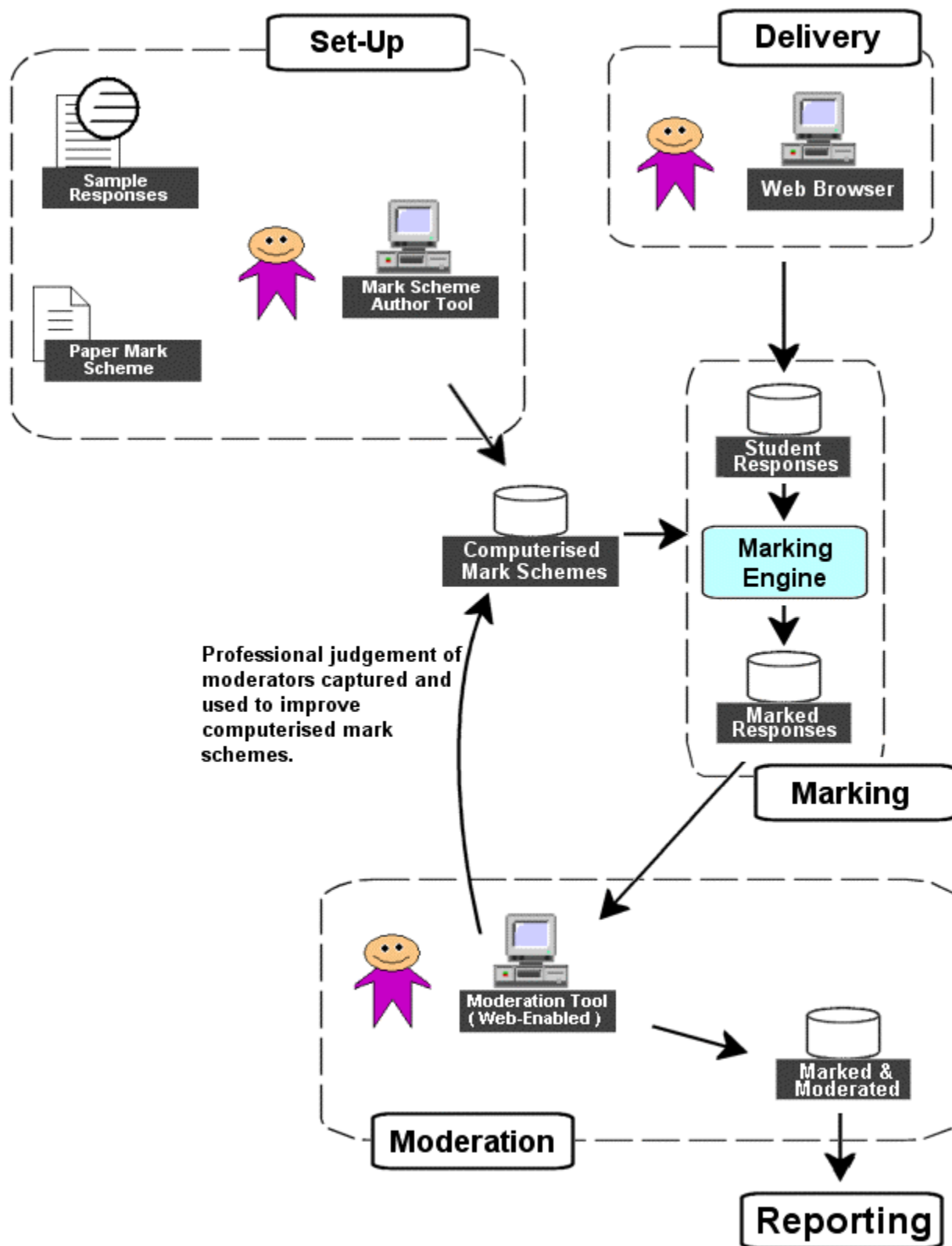
Figure 1. Overview of the Computerised Progress Test System.

Figure 2. Choosing 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.

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 3). Responses for which the moderators change the marks are highlighted (green in the web page, showing up as a grey background in Figure 3). 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 2).

UNIVERSITY OF DUNDEE

Medical School MBChB Professional Examinations

Progress Test

Moderate this question.   >>Change question   >>Home   >>Logout
>>Flag question as moderated

Orig   Code   **Question Text**
                                "A 72 year old man you are looking after is very ill and semi-comatose. You have talked to
265   GMP0JHO000095007   his wife the day before and explained things fully to her in terms of disease progression,
                                prognosis, etc. The next day his daughter telephones you from Texas and demands to know
                                all about her father's condition. What do you tell/advise her?"

View [10 ▾] responses, [correct ▾] first, then by: [Length of response (desc). ▾]   **go**

Page 6 of 17

<< previous 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 next >>

Student Answers	Mark	Change
Appologise, can not give information over the phone. She will have to speak to her mother	1	Change
tell her to call her mum for info.You have a duty to maintain confidentiality	1	Change
advice her to speak to the patient's wife first and to call back if she has any questions	1	Change
Explained to her the situation and also mentioned that her mother has been informed too.	1	Change
To discuss the situation with her mother - information can't be given over the telephone	1	Change
Advise her that her mother has all the information and suggest she talks to her mother	1	Change
Ask her to discuss the issues with the mans wife, cant really discuss over the phone	1	Change
ADVISE HER TO GET IN TOUCH WITH HER MOTHER AS SHE HAS ALREADY HAD THINGS EXPLAINED	1	Change
Can not give patient details out over the phone, suggest she phones her mother.	1	Change
Apologis but say she will have to speak with her mother about all the details	1	Change

Page 6 of 17

<< previous   1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17   next >>

Powered by Intelligent Assessment Technologies

**Figure 3. Moderating the marks for responses to an item. Marks are
amended by clicking on the "Change" link next to the relevant response.**

Outputting Results.

Subsequent to moderation, tests can be selected for output. Results are output in a flat file format suitable for subsequent processing in spreadsheets or other applications. At Dundee, the output file is processed by an application developed by Speedwell Computing Services, which provides detailed information of each student's performance categorised by content year of item, body system, curriculum outcome, and clinical block.

Running the Computerised Tests.

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. To date, year groups 2, 3 and 5 have been tested, marked, and given results, a total of approximately 460 students. Years 1 and 4 will be tested in June and July respectively. Student reaction to the computerised test was either positive or neutral – mostly students were interested only in the content of the test, not the medium.

Marking the Tests.

Year groups 2, 3 and 5 were computer marked in approximately 3 hours and 45 minutes on a 2.4GHz PC running Windows XP. This entailed marking of approximately 108,000 responses. This equates to around 30 seconds per student 'script' of 270 items.

Moderating the Computerised Tests.

The day after the Year 5 test, a moderation meeting was held to check, and where necessary amend, the marks awarded by the computer. As with the paper-based system, a group of senior academics were present, representing a broad range of expertise across the learning outcomes. The year 5 students' responses were moderated item by item, using the screens shown in Figures 2 and 3. Although the computerised system supports multiple users moderating at any one time, the decision was taken to moderate as a group as in previous years, one item at a time. To facilitate this, a computer projector was used to provide an enlarged view of the moderation screens.

To help in moderation, Intelligent Assessment Technologies (IAT) provided supplemental information on the likely performance of the computerised marking on an item-by-item basis. In effect, items were categorised according to the confidence IAT had in the computerised mark schemes – new items were lower confidence (since sample responses had not been available when configuring the computerised mark schemes), as were items where the paper-based marking guidelines were not particularly clear or explicit. The moderators targeted their moderation at the lower confidence items in the first instance, subsequently moving on to the remaining items.

The academics experience in the computer-assisted moderation process may be summarised as follows.

- Being able to view all student responses to an item together is a major advantage over the paper system, where moderation proceeded on a script-by-script basis. The variety of student responses focussed the mind on exactly what is or is not an acceptable response. Some marks were changed due to computerised marking errors, but a larger number of responses (for both "new" and "previously used" items) had their marks amended as it became clear that existing marking guidelines were too narrow, or indeed that the item was not appropriately worded. The computer-assisted moderation process made identifying such problem items far simpler than in the past.

- The academics felt that the process of moderation via computer was a largely positive experience, as opposed to the ordeal of trawling through piles of paper scripts. 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. In particular it was discovered that focussing attention on short responses (where the student may have used unforeseen abbreviations or acronyms) or longer answers (where complex sentence structures can be a problem for the current version of the marking engine (Mitchell et al 2002)) was a productive strategy.

- Although not exploited in the current project, the computer system also facilitates a distribution of the moderation responsibilities. Subjects specialists can log in to the system, and moderate items in their own specialist area. Future tests may investigate this model.

Subsequent to the moderation of year 5, Intelligent Assessment Technologies re-worked a number of the computerised mark schemes to ensure future computerised marking correlated strongly with the revised marking guidelines. The updated computerised mark schemes were then used to mark Years 2 and 3, and will later be used for Years 1 and 4.

A comparison of the processes involved in running the paper-based and computerised tests is shown in Table 1.

Paper-Based	Computerised
Test Design and Item Writing	
Same in both approaches.	Same in both approaches.
Test Preparation.	
Design, preparation, printing and collation of approximately 800 copies of the 30 page Progress Test. Multiple versions required, since tests taken on different dates have items in different orders.	(1) Tests are created by uploading an Excel spreadsheet containing the items into the test database using a simple web interface. The spreadsheet is produced during the test design process. (2) Development of the initial computerised mark schemes for 'new' questions based on the marking guidelines.
Test Delivery	
Papers are stored securely, and distributed on the day of the test. Collected after test, sorted, and stored securely.	Students sit the test in the IT suite, 80 at a time. Responses stored securely in database, and backed up.
Moderation.	
Senior academics mark the year 5 scripts and moderate the marking guidelines.	Student responses are marked by computer in batch mode after the exam (approx 30 seconds per student 'script'). The marked year 5 answers are moderated by senior academics and the marking guidelines amended where necessary.
Marking	
The moderated marking guidelines are used to mark the remaining year groups. 160 scripts per year group, a team of 6 markers can together mark around 15 scripts per hour.	The computerised mark schemes are amended in light of the year 5 moderation. The amended computerised mark schemes can then be used to mark/re-mark all year groups, 30 seconds per script.
Reporting	
The marks awarded for each individual item for each student are entered onto computer (> 180,000 items). Subsequently processed by package to produce detailed reports.	Results output in flat file format, automatically processed to produce detailed reports.

Table1. A comparison of the processes involved in the paper-based and computerised tests.

The Accuracy of Computerised Marking.

In this section the accuracy of the computerised marking is detailed.

Data from the Year 5 Moderation.

5.8% of Year 5 responses had their marks changed by the moderators. More than two-thirds of these changes arose when significant omissions or errors in the (paper-based) marking guidelines were recognised during the moderation process. In some cases changes were necessitated by inappropriate wording of an item. In others, recent changes in medical procedure or regulations required a marking guideline to be revised. In the remainder, the student responses simply highlighted inadequate marking guidelines. In all these cases, the ability to view all student responses for a given item significantly improved the ability of the moderators to focus on problems, and to revise the requirement of the marking guideline accordingly.

Leaving aside the errors due to item wording / marking guideline specification, only *1.6%* of responses had their marks changed due to erroneous computerised marking. Most of these errors were due to minor weaknesses in the computerised mark scheme (e.g. an omitted synonym), the remainder were due to system errors inherent to the marking engine. See (Mitchell et al 2002) for an explanation of these system errors.

The Re-Worked Computerised Mark Schemes.

Subsequent to the Year 5 moderation process, Intelligent Assessment Technologies were able to re-work the computerised mark schemes, 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 re-worked computerised mark schemes. The agreement between the computerised marking and the moderated marks resulting from the moderation process was *99.4%*. 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 moderation.

Validating the Computerised System.

The moderation process gave a high level of confidence in the computerised marking. As a further test however, 10 Year 2 and Year 3 students were selected at random. Their responses were hand marked using the moderated marking guidelines, and the results compared with the marks awarded by computer using the re-worked 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 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. 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 current version of 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.

- Moderation is much less painful, but also more effective using the computerised system. Academics can easily detect weaker items, with the additional advantage that collated student responses give insight into curriculum coverage.

- Items can be moderated using a subset of the cohort. After re-working of the computerised mark schemes, the remainder of the cohort can be computer marked in a few hours. This is a dramatic saving in man-power (estimated at 30 man days in the paper-based system), and a valuable improvement in turn-around time (from weeks to days). Moreover, the marking is more accurate.

- The moderated items can be re-used in future tests, with a high level of confidence in the computerised marking. As the number of items in the bank grows, fewer new items will be required, and there will be a decreasing need for moderation. It can also be envisaged that in future years, the model may change to distributed concurrent moderation by subject specialists, further streamlining the process.

- 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.

To date, year groups 2, 3 and 5 have been tested, a total of approximately 460 students. A further 300 or so remain to be tested in this academic year. Computerising the progress test has brought significant advantages to medical students and staff at Dundee, particularly in terms of eliminating the marking burden for staff, streamlining the moderation process, and providing rapid feedback to students. By utilising a state-of-the-art assessment engine, they have been able to obtain the benefits of CAA, whilst retaining the open-ended item format that they know provides highly reliable and valid tests. As the bank of moderated items grows, test administration will consume progressively less time.

Future Work.

2004 will build on the system developed and rolled-out in 2003. From Dundee's point of view, the computerised progress test will benefit from one particular technological development.

- Academics and item writers should have the ability to create and modify (i.e. moderate) the computerised mark schemes.

This functionality will require the development of a simpler interface to the marking engine. 2004 should see the introduction of such an interface.

More recent development work has provided the capability of integrating the computerised marking and moderation sub-system described here with QuestionMark Perception. Further efforts to improve this functionality will continue.

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.

ETS Technologies Website. <www.etstechnologies.com/products_c.shtml>

Fowell, S & Bligh, J. *Recent developments in assessing medical students.* Postgraduate Medical Journal. 1998; 74(867): 18-24

General Medical Council, *Tomorrows* Doctors, 1993, <www.gmc-uk.org/med_ed/tomdoc.htm>

Goldacre, M., Lambert, t., Evans, J., Turner. G. *Does medical school prepare you for medicine?*: British Medical Journal- 2003 p1011

Landauer, T. K., Dumais, S.T. (1997) *A Solution to Plato's Problem : The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge.* Psychological Review, vol. 25, pp 259-284.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). *Introduction to Latent Semantic Analysis.* Discourse Processes, 25, 259-284.

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.

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.

Veloski, J.J., Rabinowitz, H.K, Robeson, H.R., & Young, P.R. Patients don't present with five choices: An alternative to multiple-choice tests in assessing physicians' competence. Academic Medicine. 1999; 74 (5): 539-546