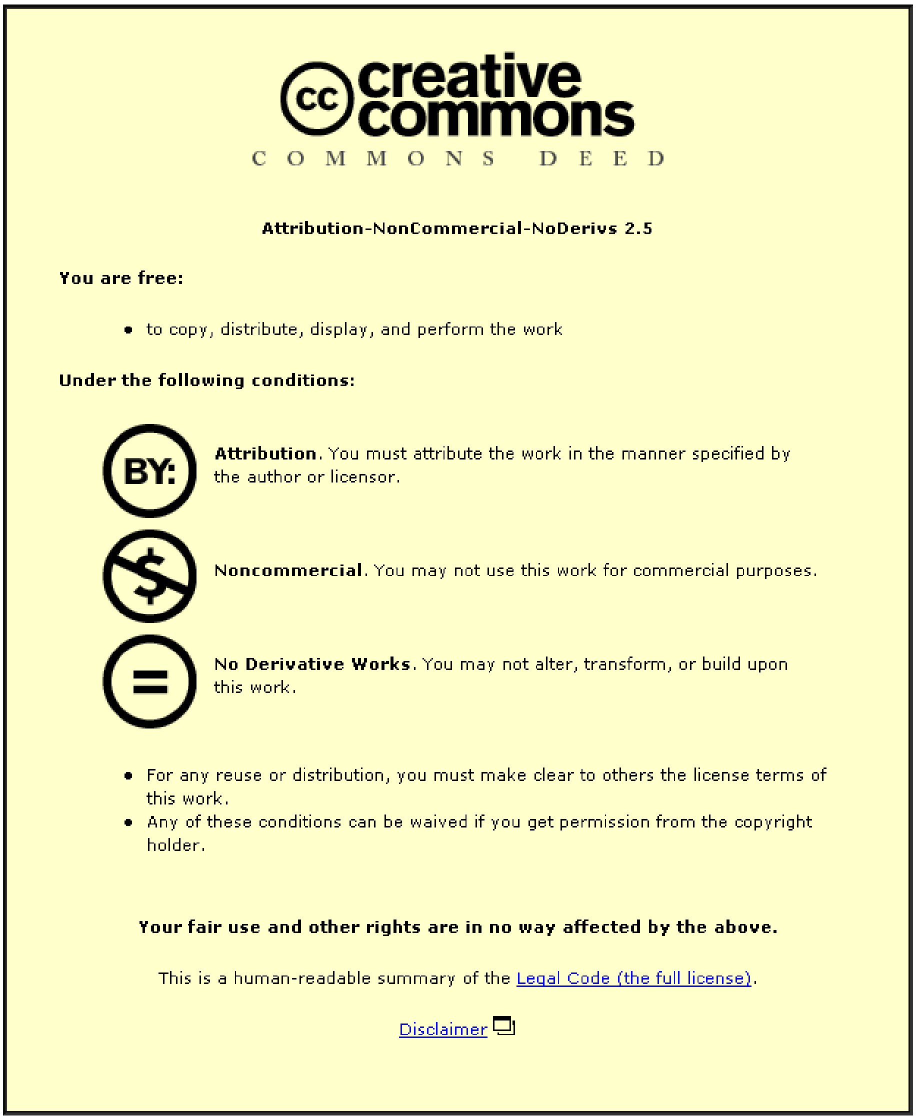


QUESTION TYPES IN ENGLISH LANGUAGE DIAGNOSTIC TESTING

Mary McGee Wood and John Morley

Question Types in English Language Diagnostic Testing

Mary McGee Wood, School of Computer Science, University of Manchester, m.m.wood@manchester.ac.uk

John Morley, University Language Centre, University of Manchester, john.morley@manchester.ac.uk

Abstract

English language proficiency testing, like large-scale testing in many other domains, often uses multiple-choice questions, to exploit the efficiency of automatic marking. An experiment supplementing established mcq tests with very short free-text answer questions in English diagnostic testing has shown that the latter are better discriminators at the lower end of student ability. Although not economic with paper-based marking on a large scale, the Assess By Computer e-assessment software offers marking options for such answers which make constructed-answer tests a realistic option.

Constructed vs selected answers in large-scale testing

In many domains, there is a need for quick and efficient large-scale testing of straightforward material. Selected answer[1] tests can be automatically marked, and are thus widely used (for example, in the UK, for the DVLA automobile Driving Theory test.)

Diagnostic testing of English language proficiency is another such domain, with global scope. Locally at the University of Manchester (UoM), the University Language Centre (ULC) tests over 1,000 students per academic year - around 850 in a single week each September - to assess their linguistic ability to follow an academic course. (This is additional to the standard TOEFL / IELTS admission requirements.)

The UoM ULC tests consist mainly of mcq's, as described below. However, selected-answer questions tightly constrain the extent to which a candidate can give evidence of incompetence, even in this intellectually limited domain. Our hypothesis was that, given the chance to answer freely, the weakest

[1] We reserve the term "objective" to mean questions to which the answer, rather than the marking judgement, is a matter of objective fact - as opposed to "subjective". Objective questions can require constructed answers, and subjective questions selected answers ("Give your opinion on a scale from 1 to 5 …")

candidates would give evidence of greater weakness than could be seen from MCQ results. This was confirmed by the data.

English language diagnostic testing

The test employed by UoM ULC for English language diagnostic testing is the Chaplen Speeded Grammar and Vocabulary Test (Chaplen, 1970), which has been used at Manchester University since the early 1970s. It is a well tried and tested gauge of a learner’s knowledge of the English language system and its formal or “educated” vocabulary. The total number of correct answers is presented as a percentage score. The test discriminates well at the upper intermediate and advanced levels of language proficiency, with students at these levels typically attaining scores ranging from 50% - 90%. A score of more than 90% indicates that the learner is approaching native speaker level. Below intermediate levels (approx 40%), however, it is not a useful instrument as it begins to lose its discriminatory power. High marks on the Chaplen typically correlate well with the number of years studying English as a foreign language in formal settings, though the strength of this relationship has not been tested.

Originally developed in the late 1960s, the test reflects the structuralist description of language and methods of language testing. Using a multiple choice format, the Grammar (10 mins) and Vocabulary (18 mins) sections test students' knowledge of a range of individual items of structure and lexis in “everyday educated English” (Chaplen, 1970: 174). For each section, there are 50 questions, each consisting of a sentence with a word or phrase omitted. The test taker must choose the correct filler from the list of possible answers provided. There is a choice of three possible answers in the Grammar section and five possible answers in the Vocabulary section. The short amount of time allowed for each section means that students work under considerable pressure of time and only the more proficient students manage to complete all the questions. The test is quick to administer and quick to mark. This is one of its major advantages, since it permits the rapid processing of very large numbers of students at low cost. Combining this rapidity of administration with an OMR marking system means that up to 1000 students can be tested and given their mark within a few days.

The theoretical assumption which underpins the use of the test at Manchester is that adequate knowledge of the general language system can serve as a reliable indication of a student's ability to apply this knowledge in academic situations. Its principal use is to identify recently arrived overseas students who would benefit from attending classes in academic writing provided by the University, or who will probably experience difficulties in their academic work due to less than adequate levels of English language proficiency in reading and writing. A score 40% or less, broadly indicates that a student has an inadequate level of English language proficiency for academic study. The extensive trials that the test underwent during its development would appear to support this (Chaplen, 1970). In addition, in two follow-up studies, the test has been shown to have reasonable predictive validity (James, 1980; O'Brien, 1993).

Because Chaplen is basically a test of discrete item recognition, it has to be complemented by a piece of continuous writing. The writing test consists of three questions to which short "essay" answers are expected, to be completed in 30 minutes. Morley (2000), who made a number of improvements to the test, has shown that the test scores correlate quite strongly with assessment of students' continuous writing using trained assessors[2].

Despite all its advantages, it needs to be emphasised that Chaplen is not a test of language production, and it is not a test of language skills. In fact, it assesses a fairly narrow aspect of language competence through the recognition of correct lexical and grammatical choices provided as part of an artificially restricted set of choices. In this sense, it is less of a finely tuned instrument than the much more sophisticated, and much more expensive, internationally recognised university entry tests (eg IELTS and TOEFL) which take very much longer but which also test a broad range of language skills.

Furthermore, despite the strong correlations with writing scores mentioned above, it is still not uncommon to come across cases of good spoken and written communicators who do not score well on Chaplen, and of good Chaplen scorers who are not good communicators. Finally, because of the multiple choice design, the discriminatory power of the test below a certain level is weak (around 40%) and even non–existent (around 25%). We therefore sought ways of maintaining the efficiency of the instrument, whilst at the same time endeavoring to measure students' ability to *produce* correct language rather then simply to choose it. The aim was to fine-tune the instrument and to increase its discriminatory power without any loss of efficiency.

The experimental tests

Our hypothesis was that, if a practical way could be found of testing with free-text questions - even with single-word answers - (a) all the students would be more effectively challenged by what would become, in effect, a production rather than a recognition task; (b) the weakest students would make more extreme errors than any of the mcq distractors, and we would thus have more effective discrimination at the bottom end of the range. The ABC (Assess By Computer) e-assessment software (Sargeant et al 2004) developed at UoM looked promising, and has been used in (to date) two trial runs of free-text question tests, with a third scheduled.

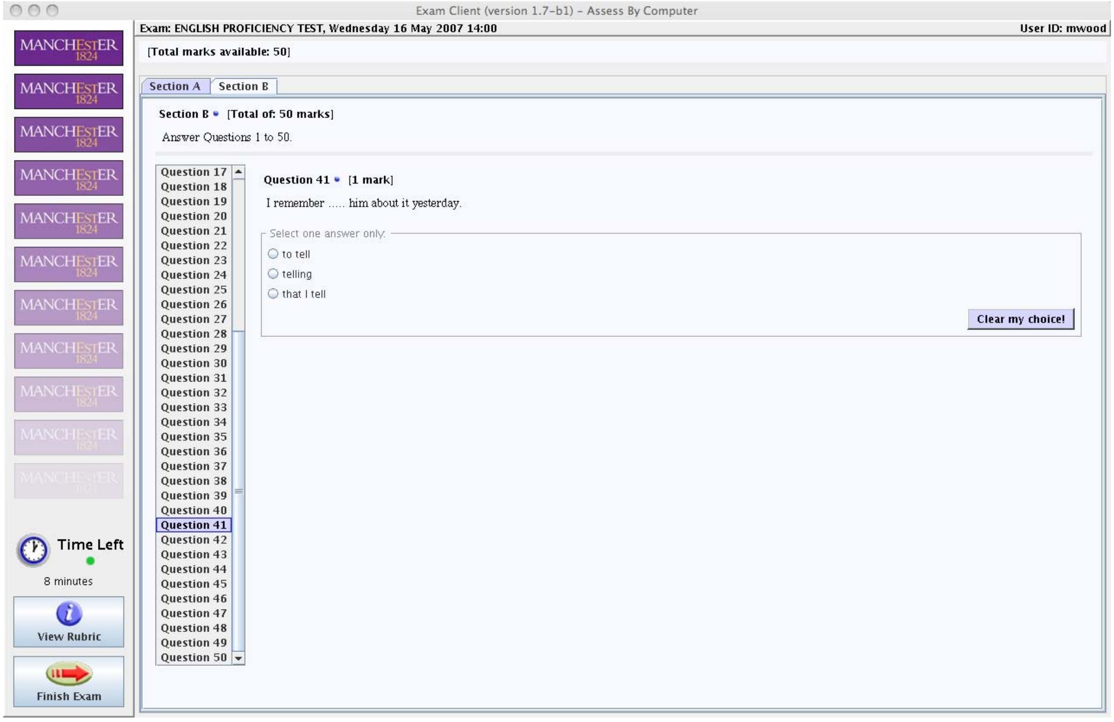
The original UoM English language proficiency diagnostic assessment, as described above, consists of three separately timed tests[3]: "Grammar and

[2] Spearman ranking correlations between Chaplen subtests and writing: grammar .696; vocabulary .809 (no. of cases 153; correlations significant at p < .001 1-tailed).

[3] The terminology we use here is as follows:

Usage" (10 minutes), "Vocabulary" (18 minutes), and "Writing" (30 minutes). These tests were set up in the ABC software and first taken in this form in February 2007 by 23 students (January entrants to postgraduate programs in the School of Computer Science, UoM).

Although the students were unfamiliar with the software, none showed any signs of difficulty in using it, and results were as expected from previous experience with similar groups, i.e. there was no evidence of bias caused by use of the software. The clear difference lay in the speed of marking. The MCQ tests were marked automatically, with marking complete within minutes of the last student submitting their answers, rather than waiting several days for a scanning service. The answers to the "writing" test were output as a pdf file and marked on paper: the saving here lay in the greater ease of reading typescript than handwriting.



Exam Client (version 1.7-b1) – Assess By Computer

Exam: ENGLISH PROFICIENCY TEST, Wednesday 16 May 2007 14:00 User ID: mwood

[Total marks available: 50]

Section B • [Total of: 50 marks]

Answer Questions 1 to 50.

Question 41 • [1 mark]

I remember ..... him about it yesterday.

Select one answer only:

○ to tell

○ telling

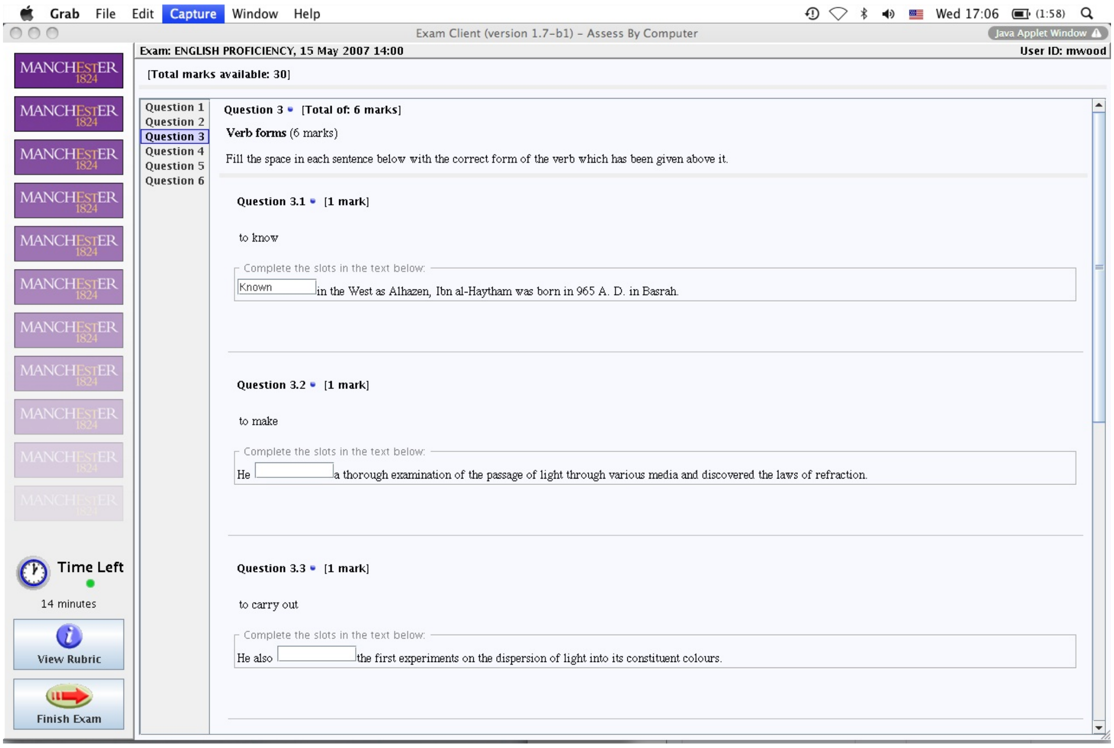
○ that I tell

Fig. 1: The ABC exam client, showing a multiple choice question

Thus re-assured, we proceeded to supplement these established mcq and writing tests with an experimental grammar test using free text entry rather than multiple choice questions, i.e. constructed rather than selected answers. This "Grammar2" test comprised five questions, each with six leaves, to be answered in 15 minutes. The leaves in the first four questions each took the form of a sentence with a slot to be filled in with a word of a specified

Assessment: an exam taken in one session, at one sitting; it may comprise multiple *tests*.
Test: a discrete section of an assessment which can have a distinct time constraint; it will normally comprise multiple *questions*.
Question: as conventionally used: the highest level of division within a test; it may ultimately comprise multiple *leaves*, possibly with intermediate levels of tree-structured sub-questions.
Leaf: the smallest discrete unit to which marks can be assigned.

category: determiners and pronouns, verb forms, prepositions, and modal verbs. In the final question, the students were asked to place the words from a given list in the correct order to form a grammatical sentence. By running all three forms of test in parallel, we were able to compare the performance of each student across the formats, and to begin to assess the effectiveness of the free text test. And of course, the new gapfill and sentence order items allow us to test production in terms of word choice (gapfill) and syntax (reorder items). So these new test items go some way towards overcoming the limitations of Chaplen.



Exam: ENGLISH PROFICIENCY, 15 May 2007 14:00

User ID: mwood

[Total marks available: 30]

Question 3 [Total of: 6 marks]

Verb forms (6 marks)

Fill the space in each sentence below with the correct form of the verb which has been given above it.

Question 3.1 [1 mark]

to know

Complete the slots in the text below:

Known in the West as Alhazen, Ibn al-Haytham was born in 965 A. D. in Basrah.

Question 3.2 [1 mark]

to make

Complete the slots in the text below:

He ______ a thorough examination of the passage of light through various media and discovered the laws of refraction.

Question 3.3 [1 mark]

to carry out

Complete the slots in the text below:

He also ______ the first experiments on the dispersion of light into its constituent colours.

Fig. 2: The ABC exam client, showing free text slot questions

To illustrate the comparisons between the grammar question formats, these examples test the students' knowledge of English prepositions, verb forms, and modal verbs respectively:

He solved the problem ..... two minutes.

A. at   B. in   C. with

These factors allowed computers to be produced ______ an unprecedented commercial scale.

I remember ..... him about it yesterday.

A. to tell   B. telling   C. that I tell

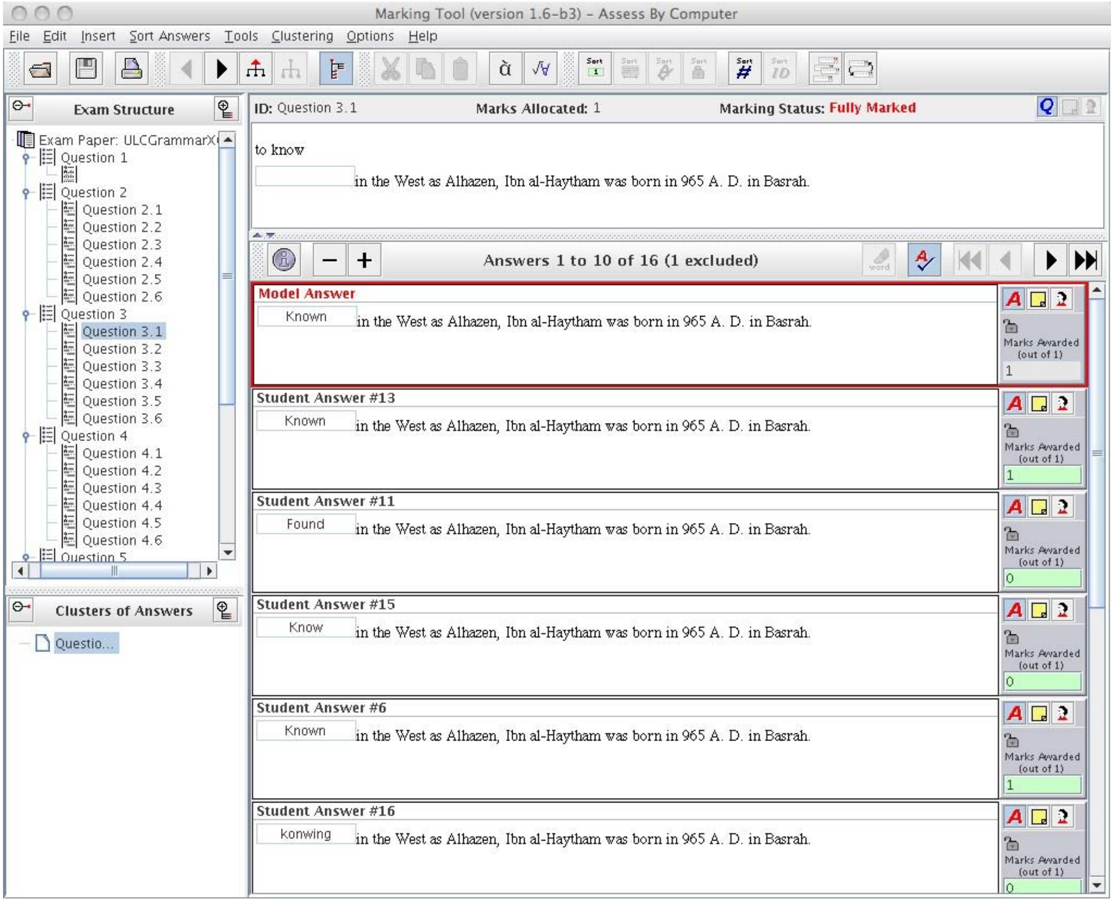
(to know) _____ in the West as Alhazen, Ibn al-Haytham was born in 965 A. D. in Basrah.

```
We'd get there on time if we ..... a taxi.
    A. could find     B. able to find     C. can to find
My sister _____ speak three languages by the time she was 12.
```

The four-test assessment, including the experimental constructed-answer (gapfill) form of the grammar test, was first run in May 2007 for 17 students in the UoM School of Mathematics: these students had been identified as having language difficulties, so the expected standard was low. The data used here is drawn from that assessment. A second run, in September 2007, for 56 entrants to taught postgraduate students in the UoM School of Computer Science, revealed similar patterns.

Marking techniques

The ABC marking software offers several options which are particularly useful in the marking of very short free text answers, including various highlighting and sorting options as well as (in some cases) automatic marking.



ID: Question 3.1 Marks Allocated: 1 Marking Status: Fully Marked

to know

in the West as Alhazen, Ibn al-Haytham was born in 965 A. D. in Basrah.

Answers 1 to 10 of 16 (1 excluded)

Model Answer
Known in the West as Alhazen, Ibn al-Haytham was born in 965 A. D. in Basrah.
Marks Awarded (out of 1) 1

Student Answer #13
Known in the West as Alhazen, Ibn al-Haytham was born in 965 A. D. in Basrah.
Marks Awarded (out of 1) 1

Student Answer #11
Found in the West as Alhazen, Ibn al-Haytham was born in 965 A. D. in Basrah.
Marks Awarded (out of 1) 0

Student Answer #15
Know in the West as Alhazen, Ibn al-Haytham was born in 965 A. D. in Basrah.
Marks Awarded (out of 1) 0

Student Answer #6
Known in the West as Alhazen, Ibn al-Haytham was born in 965 A. D. in Basrah.
Marks Awarded (out of 1) 1

Student Answer #16
konwing in the West as Alhazen, Ibn al-Haytham was born in 965 A. D. in Basrah.
Marks Awarded (out of 1) 0

Fig. 3: The ABC marking tool, showing free text slot questions

Slot questions are shown in Figs. 2 and 3: a string is displayed which includes an in-line empty box into which the answer is typed. Slot questions can (but need not) be marked automatically, with exact matches to the model answer awarded full marks and anything else none. Where appropriate, this is the fastest way of marking a set of student answers if one is certain there is only one right answer - so that the automatic marking can be trusted - and still highly efficient if one is prepared to spend some time hand-checking the “wrong” ones and possibly over-riding the automatically assigned mark in the event of discovering an unexpected alternative correct answer.

Dynamic marking, when enabled, will automatically propagate a mark manually assigned to one student answer to any others which are identical text strings. This is a highly effective way of cutting through a large student answer set which contains a high proportion of identical answers (e.g. where 150 out of 350 Life Sciences students answer a question with “Blood group A”). The software can assign marks automatically, but only in highly conservative and constrained circumstances and under human control.

Cluster by keyword, given a specified text string (word or phrase), will group together all the student answers which contain it, and highlight it within them. Keywords can be set up in advance - typically based on the model answer - or added dynamically, “on the fly”, in response to patterns observed in the student data. This process does not suggest marks: that is left to the human marker, who will however find both speed and consistency of marking greatly improved compared to paper-based marking.

Cluster by similarity (Wood et al 2006) - this is at its most effective applied to very short answers such as we have here. It again does not suggest marks. Its advantages over keyword clustering are that it does not require keywords to be set in advance, and that it produces a count of the number of items in each cluster, thus automatically quantifying the mistakes. Its main drawback for language testing is that the default pre-processing which optimises its performance on longer answers (10-30 words) removes some of the information which is actually critical here, such as inflexional endings, word order, and the presence or absence of small, semantically insignificant words.

Thus the stripping of affixes (stemming) which allows the system to recognise the conceptual similarity of “penguin” and “penguins” will also ignore the difference between “carried” (correct) and “carries” (incorrect). And in an example from a Biology test (reported in Wood et al 2006), 11 distinct variants on

> *haemoglobin concentration in the blood*
> *Haemoglobin Concentration of the blood*
> *Concentration of haemoglobin in blood*
> *Concentration of haemoglobin in the blood*

etc, were usefully grouped together by similarity clustering. Clearly, however, this would make nonsense of a word-ordering question! We are looking at ways of tuning the similarity clustering to language testing, but for now it remains inappropriate.

We chose to set up the ULC Grammar2 “gapfill” test using slot questions with automatic marking.

Results

As expected, the range of incorrect student answers in the constructed-answer test was far wider that of the distractors offered in the selected-answer version, both in number and severity. Taking the examples shown earlier:

He solved the problem ..... two minutes.

A. at B. in C. with

A - 0

B - 14

C - 3

These factors allowed computers to be produced ______ an unprecedented commercial scale.

by (6)
in (5)
as (1)
because (1)
out (1)
to (1)
on (0)

I remember ..... him about it yesterday.

A. to tell B. telling C. that I tell

A - 9

B - 5

C - 3

(to know) _____ in the West as Alhazen, Ibn al-Haytham was born in 965 A. D. in Basrah.

known (6)
kown (1)
knowing (3)
konwing (1)
knows (2)
to know (1)
knowledge (1)
found (1)

We'd get there on time if we ..... a taxi.

A. could find B. able to find C. can to find

A - 12

B - 5

C - 0

My sister _____ speak three languages by the time she was 12.

could (8)
can (6)
has been able to (1)
was (1)
she (1)

This (small but representative) sample of our student data does show that, while some students made mistakes of the sort which might well be found as distractors in a sensibly designed MCQ, others were far worse. Some revealed a misunderstanding of the question leaf, such as

By the 1980s, computers had become sufficiently small and cheap to replace simple mechanical controls _____ domestic appliances such as washing machines.

where only one student gave the correct answer “in”, while six said “by” and two “with”. The most egregious were those which revealed a failure to understand the overall question, such as using a word which was not of the part of speech requested.

Prepositions:

Vacuum tube-based computers were ______ use throughout the 1950s.

widely (3)
to (3)
being (3)
in (2)
be (1)
into (1)
introduced (1)
introdued (1)

Verb forms: *knowledge, translation*
Modal verbs: *she, already, only, able, fill, got, type, tip*
Determiners and pronouns: *with, to, of, by, on, in , what, people, astronomer, about, at*

Future experiments will include negative marking for the worst errors, formalising and quantifying (albeit with an element of subjective judgement) these indicators of student ability.

Conclusions

Simply moving the original UoM ULC diagnostic tests into the ABC software brought immediate significant benefits in speed and convenience of administration for moderately sized groups (c10 – 90) where there is access to a computer cluster. The new experimental free-text answer test has added greater discriminatory power with no necessary increase in marking time, although it also adds the option of scrutinising the student answers marked as incorrect if desired. A typical *modus operandi* would be to accept the automatic marking judgements initially - given the time pressures involved at the time - and go back at leisure to analyse the student data.

An important incidental benefit of ABC assessment has always been the ability to “mine” student answer data for a wealth of information about both students and assessments (see e.g. Wood et al 2005). Most obviously, common errors are easily identified, such as

You _____ to have included this source in the reference list.

ought: 1
have: 6
must: 5
should: 2
would: 1
able: 1
ask: 1

Alhazen (to be) ______ featured on the obverse of the Iraqi 10,000 dinars banknote issued in 2003. The asteroid 59239 Alhazen was also named in his honour.

is: 0
was: 9
has been: 2
had been: 1
was been: 1
been: 2
being: 1
-- : 1

Such clusters are significant paedogogically, as they indicate either a genuine common misunderstanding which should be addressed in teaching (as in the first example), or a weakness in the question, which should perhaps be reviewed before being used again (as in the second example, where “was” in the second sentence makes “was” an attractive option in the first). Although calculation of the discriminatory power of mcq’s is commonplace (IML 2007), this goes further, by revealing weaknesses which a question setter - for a number of reasons - would probably not consider using as distractors.

The only significant limitation that we can see on e-assessment of English language proficiency using short-free-text questions is the availability of computer infrastructure. As noted above, over 1,000 international students take the UoM ULC’s English Language Diagnostic tests each year, the great majority of them during registration week each September. There are simply nowhere near enough machines in suitable locations to allow simultaneous testing on this scale. This prevents the ABC assessment system from replacing the original closed-item multiple choice format using OMR marking for large groups (although the capacity of the software is more than adequate).

Wider prospects

The results of our experiments so far indicate that our original hypothesis is correct: that, at least for English language diagnostic testing and at the lower end of the ability range, even single-word constructed answer questions can be better discriminators of student ability than selected-answer questions.

Diagnostic testing does have some distinctive aspects as compared with formative or summative assessment. All that is needed is a rating of students' ability adequate to determine (in this case) whether further English language study is needed. Criticality is low, and conservatism in marking appropriate: the worst that can happen is that a student can be recommended to take a course which s/he does not really need, or not to take a course which s/he does need, of which the former is of course the lesser evil.

However our findings have wider implications for the use of selected-answer questions in general. Such questions are widely used for efficiency reasons, and paedogogic defences have been offered, but are unconvincing: "It is claimed that skilled items writers can develop items to test higher level intellectual skills (Cannon and Newble: 1983) but if the perception of students is that these types of questions usually test the recall of facts, then they will prepare for them accordingly." (IML 2007) Parallel pilot studies in other areas currently dominated by selected-answer testing may well open the way to practical and efficient, but more flexible, challenging, and effective e-assessment.

References

Cannon, R.A. and Newble, D. (1983). A Handbook for Clinical Teachers, Lancaster, MTP Boston: p 97-105.

Chaplen, E. F. (1970). *The identification of non-native speakers of English likely to underachieve in University courses through inadequate command of the language.* Unpublished PhD Dissertation, University of Manchester.

IML 2007. *Types of assessment*. http://www.iml.uts.edu.au/assessment/types/mcq/index.html (Originally published in Trigwell, K. (1992). Information for UTS staff on Assessment. Sydney: UTS Working Party on Assessment). University of Technology Sydney, Institute for Interactive Media and Learning.

James, K. (1980). *Survey of University of Manchester Overseas Postgraduate students' initial level of competence in English and their subsequent academic performance: Calendar year 1977*. **ELT Documents 109 - Study Modes and Academic Development of Overseas Students,** G. M. Greenall and J. E. Price, eds., The British Council.

Morley, J. (2000). *The Chaplen test revisited*. **Assessing English for Academic Purposes**, G. Blue, J. Milton, and J. Saville, eds., Peter Lang, Oxford, 49-62.

O'Brien, J. P. (1993). *English for Academic Purposes: The Role of the Subject Tutor.* NATESOL, The University of Manchester.

Sargeant, J., M.M. Wood & S. Anderson. (2004). *A human-computer collaborative approach to the marking of free text answers.* 8th International Conference on Computer Aided Assessment, Loughborough, UK. pp. 361-370.

Wood, M.M., C. Jones , J. Sargeant & P. Reed. (2006). *Light-weight clustering techniques for short text answers in HCC CAA.* 10th International Conference on Computer Aided Assessment, Loughborough, UK.

Wood, M.M., J. Sargeant & C. Jones. (2005). *What Students Really Say.* 9th International Conference on Computer Aided Assessment, Loughborough, UK. pp. 317-327.