Title	The Use of Computers in English Language Testing.
Author(s)	Michael Vallance
Source	*Teaching and Learning, 24*(1), 67-76
Published by	Institute of Education (Singapore)

Teaching and Learning, Vol. 24, No. 1 (June 2003) pp. 67–76

The Use of Computers in English Language Testing

Michael Vallance

Abstract

The aim of this paper is to report on the use of computers in high and low stakes English language testing. The paper starts with an outline of the functions of English language testing in order to contextualise the electronic varieties. An explanation of Computer Adaptive Testing in high stakes testing is then provided. Computer use in low stakes testing is demonstrated via personally designed examples.

Functions of English Language Tests

The functions of an English test are to measure a learner's language proficiency, to diagnose a learner's strengths and weaknesses, or to assist in the placement of the learner on an appropriate language course.

English tests are undertaken prior, during or at the end of a course. For instance, a programme can be altered to better suit the needs of the learners based upon test scores during a course of study. This is called formative testing. Then, at the end of the programme, a final test can be administered. This is called summative testing.

English tests can be objective tests where candidates are scored against established responses, as is often the case in multiple-choice tests. On the other hand, subjective tests may be exemplified by free composition essays, which may then be objectified through precise ratings identifying, for example, types of errors to be quantified.

Although language learning theories and teaching methods have evolved over the past 40 years (from behaviourist-inspired Audiolingualism to fluency-driven Communicative Language Teaching), the parallel development of computer-based testing has been disappointing. Today's high speed networked computers test English language skills in the same manner as the mainframe servers of the 1960s.

High and Low Stakes English Language Testing

High stakes testing is the term given to tests that have a central role in determining who will and who will not gain access to employment, education or certificates of accreditation. A familiar high stakes test for English language learners is TOEFL (a Test of English as a Foreign Language), which accredits a foreign student's language level in order to determine whether that student has crossed a university entry level threshold. However, TOEFL primarily tests reading and listening comprehension. Written discourse is not always tested and a speaking test (pronunciation, prosodic features or comprehensibility) is rare.

Another high stakes testing example currently gaining in stature in Singapore is the Scholastic Aptitude Test (SAT) where scores determine whether a student qualifies for course entry to a university or junior college. The English component of the SAT measures a student's knowledge of the meaning of words, the ability to see a relationship between a pair of words, and how the parts of a sentence logically fit together. Questions are primarily multiple-choice (MCQ) and sentence completion formats.

Research has shown that such well-developed tests in the knowledge, skill and ability domains are indeed valid for their intended purpose and are useful predictions for future performance in employment and academic settings (Sackett, 2001). However, there are many reports and web sites that refute such claims (Amrein & Berliner, 2002).

Low stakes tests generally do not carry the trepidation associated with the formal accreditation of high stakes testing exemplified above. An example can be a typical secondary class of pupils requiring individual help with reading comprehension, grammatical structure or vocabulary awareness. The teacher can become a low stakes tester. Traditionally she would develop paper-based worksheets, which would probably be marked late into the night, giving correct/incorrect scores and often no feedback. However, careful utilisation of computer-based testing coupled with an informed use of the computers can add value to a language lesson, as exemplified later in this article.

The next section will exemplify the use of computers in high and low stakes English language testing.

High Stakes Computer-based Testing (CBT)

The growth of state-of-the-art computer enhanced English language learning environments has been impressive. The development from the PLATO (Programmed Logic for Automatic Teaching Operations) network in the 1960s, where it was possible for instructors with no technical or programming skills to write exercises for their students, to the current use of synchronous (e.g. Internet Relay Chat and Multiple Object Oriented environments) and asynchronous tools (e.g. e-mail and bulletin boards) has been well documented (Higgins & Johns, 1984; Warschauer, 1996).

Despite the use of technology in second language learning during the 1960s and early 1970s, the first clear evidence that CBT was taken seriously by academia was in 1975 when the University of Iowa held the first Language Testing and Research Conference. The conference dealt with latent trait theory (models of probability used to score a learner's ability and a test item's difficulty), and test item bank construction and selection. This may seem rather late to our science colleagues but the English language teachers were promoting performance-based testing such as group interviews and role plays which reflected the Communicative Language Teaching approach prevalent at that time. Such testing was not readily suited to CBT formats.

In 1991 Madsen reported on a Brigham Young University computer based foreign language placement test. The listening and reading comprehension tests were reported to offer the following benefits:

> *... on average only one-third as many items needed to be presented to students as on a conventional ESL tests in use at the university and students required only one-third to one-fourth of the time. Over 80% of the examinees responding to the affect questionnaire found it less stressful to take a computer adaptive language test than conventional English language tests. Moreover, lack of computer experience was not found to be related to degree of anxiety while taking the computer test (Madsen, 1991; 237–238).*

This electronic test was known as a Computer Adaptive Test (CAT); a test that adapts itself to individual test takers (see below). When comparing the pencil-and-paper results to the CAT results it was found that the "CAT assessment tends to generate higher working levels of performance" (Stevenson & Gross, 1991: 230). A number of advantages of CAT were thus stated:

- by predetermining degree of test difficulty and cut-off scores, the CAT was targeted to specific ability levels of the students;
- practice effects were minimised due to the random selection of test items;
- test administrators were provided with immediate feedback of items answered incorrectly thereby allowing for diagnosis of learning problems;
- immediate feedback was provided to the student;
- students could work at their own pace;
- time saving; from 2 hours per traditional test to 18 minutes for an equivalent CAT.

This was seen as a major breakthrough in legitimising CAT. A number of institutes later developed their own versions of CAT such as HyperCAT, DIALANG and CommuniCAT, to run on proprietary hardware and operating systems.

Computer Adaptive Testing

A CAT is a test that adapts itself to individual test takers on the basis of previously selected answers to given questions (known as items). For example, to begin a test,

items are retrieved from a large computer database. If a student answers an item correct he will then be presented with another item of greater difficulty. If incorrect, an easier item is offered. This is repeated until an optimum level is established: point < > in Fig. 1. With reference to Fig. 1, the end of success is at point + in which the student last succeeded with three successive items. The beginning of failure is at point – where three successive fails were recorded. With items pre-calibrated and arranged for difficulty, the ability estimate continues in this iterative manner until the optimum level is established.

A CAT utilises latent trait measurement where an item difficulty scale is independent of the ability differences of any particular sample of examinees; called Rasch one parameter content trait measurement. Briefly, candidates and items are judged according to the likelihood of their response patterns given the observed ability and item difficulty. Complex formulae, outside the scope of this paper, are used to determine this. Latent trait measurement offers a number of advantages over traditional testing and are summarised in Table 1.

However, there are a number of disadvantages to consider. Students must answer each question to advance. They cannot skip questions or answer the easier questions first. Also, despite claims of cost savings, a CAT is expensive. For instance, the commercially operated English Testing Services (ETS) surprisingly announced on its website in May 2002 that it was closing 84 test centres and reverting to paper-based testing. Furthermore, an ABC News journalist reported that a PhD candidate's score decreased from 690 on a paper-based test to 300 when tested using CAT; at which point she would have been refused entry to a law course for which she was highly qualified. Finally, it has been reported that the use of computer monitors encouraged inefficient or counter-productive reading behaviour in language learning (Windeatt, 1986).

Despite its criticisms CAT is still used by ETS for measuring a learner's language proficiency, and diagnosing his strengths and weaknesses. The technical and mathematical expertise required for the development of a CAT is beyond the expertise of English language teachers. However, commercially produced testing templates are available from ETS for use as formative and summative tests.

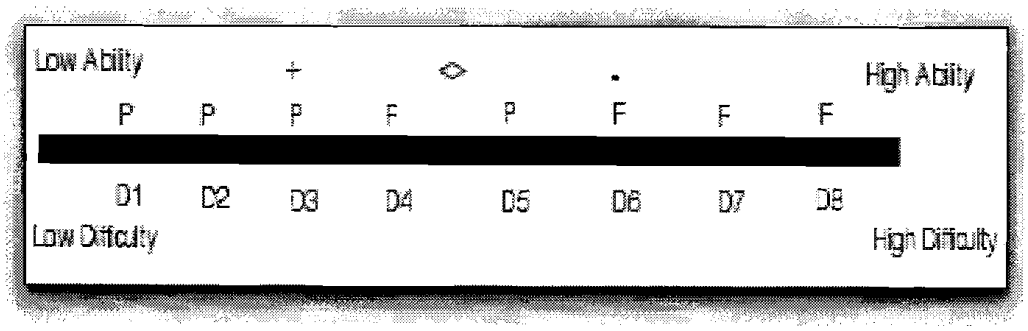

Fig. 1. Illustration of the computer adaptive testing procedure.

Table 1.
A comparison of traditional testing and latent trait measurement.

Traditional testing	Latent trait measurement
It is not possible to administer one test of, for example, reading comprehension A to student X and reading comprehension B to student Y, and make direct comparisons; unless a large random sample take both tests.	It is possible to compare abilities using different tests by referring to a link of common items. Items are calibrated and joined to form a common bank of items and any cluster of these items may be used to measure ability on the same scale.
The measurement of ability tends to be more reliable near the mean of the distribution rather than at its ends.	A standard error of measurement is determined for every possible point. Thus reliability goes beyond a global estimate for a given test but is associated with every possible candidate and item on that test.
There is no attempt to differentiate between wrong responses that are truly incorrect or simply guesswork.	Testers can quantify the improbability of any response given the knowledge of the difficulty of the item and the ability of the candidate. Once items have been calibrated for difficulty it is possible to select items to match the known range of the candidate.

The paper will now look at low stakes English language computer tests which can be produced by English language teachers.

Low Stakes Computer-based Testing

Hot Potatoes

Hot Potatoes is software that allows the teacher to create online MCQ, cloze, text matching, and word-order tests through a user-friendly interface. Word and letter clues are offered upon request and the online dictionary can be easily accessed for help. Digitised video clips may be included for contextualisation and related comprehension questions offered. The program also enables a teacher to input feedback for a student who selects an incorrect response. This may be useful as formative assessment to help improve student performance via the teacher-constructed feedback. A final score is then given, which can be viewed equally by the student and teacher (see Fig. 2). The student consequently develops the skills of an independent learner while the teacher's role is that of a steward; guiding, facilitating, managing, monitoring and helping learners in her care. Hot Potatoes is available online at http://web.uvic.ca/hrd/halfbaked/.

TexToys

Text reconstruction software, popular with language learners in the 1980s, is now available in a browser format for access online. The students have to complete

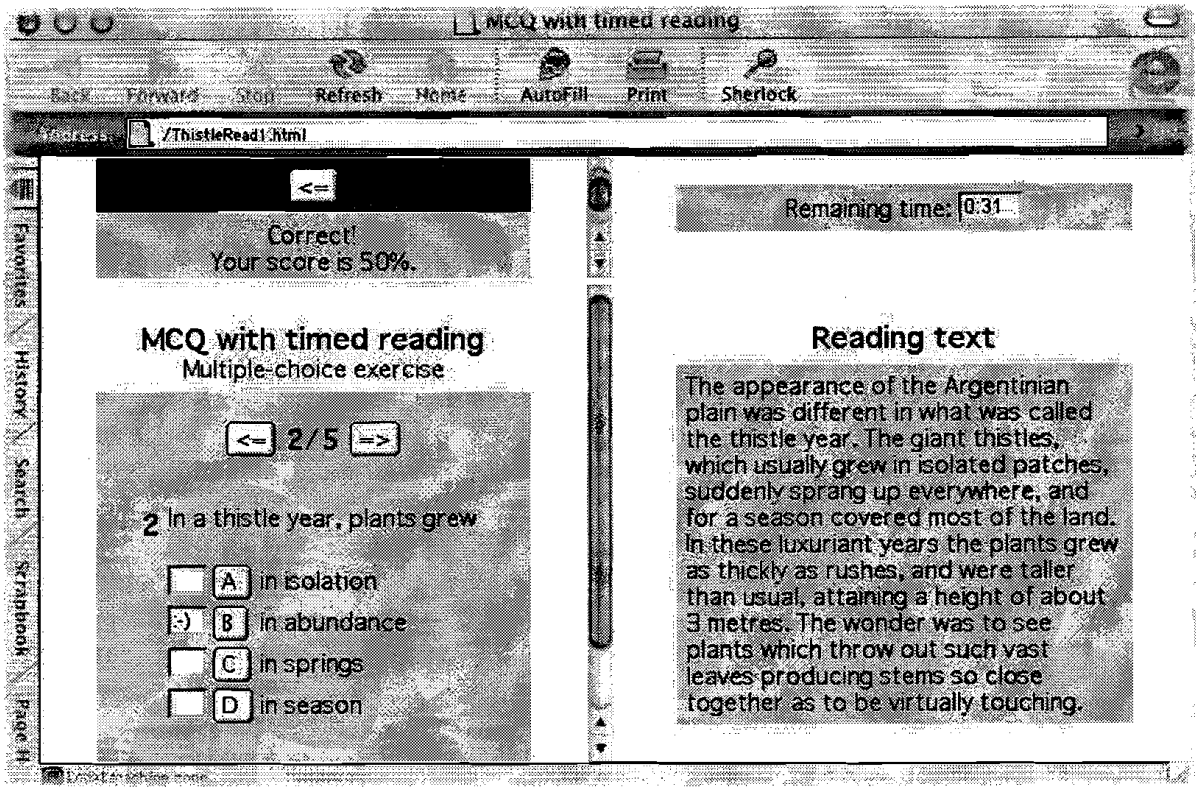

Fig. 2. An online timed reading test with MCQ.

a text by inserting missing words inferred by the surrounding text clues (see Fig. 3). Again the interface is user-friendly and teachers with limited computer experience can quickly design language exercises. TexToys is available online at http://www.cict.co.uk/software/textoys/index.htm.

Web-based Reading Mazes

Web-based reading mazes, originally designed in paper-based format by Rinvolucri, lend themselves to testing students' knowledge of notional and functional language. For example, Business Meetings combines functional–notional language and Rinvolucri's reading maze task via an online role-play (see Fig. 4). Business Meetings is online at http://www.celt.stir.ac.uk/staff/higdox/vallance/diss/fp.htm.

Randall's Cyber Listening Lab

Low stakes listening tests for English language learners are also available online in digital format. Randall's Cyber Listening Lab has extensive listening files with well constructed pre- and post-listening comprehension questions. Learners can select a level of difficulty and an interesting context. At the end of the test the participant is immediately provided with a score. Randall's Cyber Listening Lab is online at http://www.esl-lab.com/.

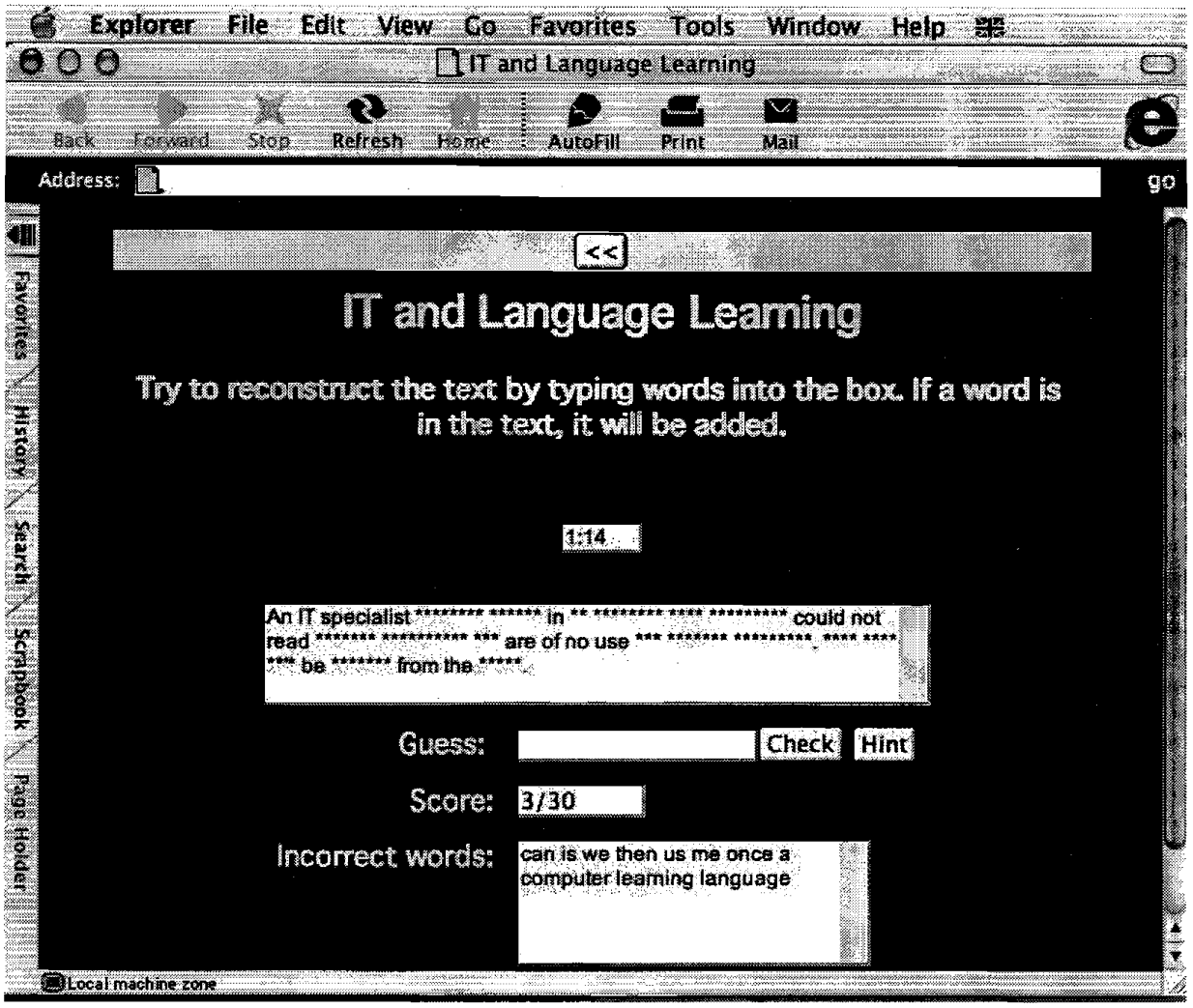

Fig. 3. An example of a text reconstruction test.

BlackBoard

Similar to the PLATO network of the 1960s (mentioned above) another web-based system for electronic learning is BlackBoard. This system can store documents for students' retrieval, add announcements, insert video clips, create bulletin boards (BBS), and can authenticate and track users. It also has a test function where teachers can create tests such as MCQ or open-ended response questions via dialog boxes. Questions are stored in a bank for test customisation, although item selection has to be done manually. The tests may be utilised as formative tests where the teacher can provide group or personalised feedback either during face-to-face tutorials or on the distance-based electronic BBS. Due to the security features of BlackBoard it is feasible to facilitate a test in a closed computer lab under strict supervision. This may be considered an example of low stakes testing where marks achieved would form *part* of an overall course grade. BlackBoard is available at http://www.blackboard.com.

A similar example is that offered in Macromedia's Flash format by World Class Arena where users are authenticated locally by an administrator but their results

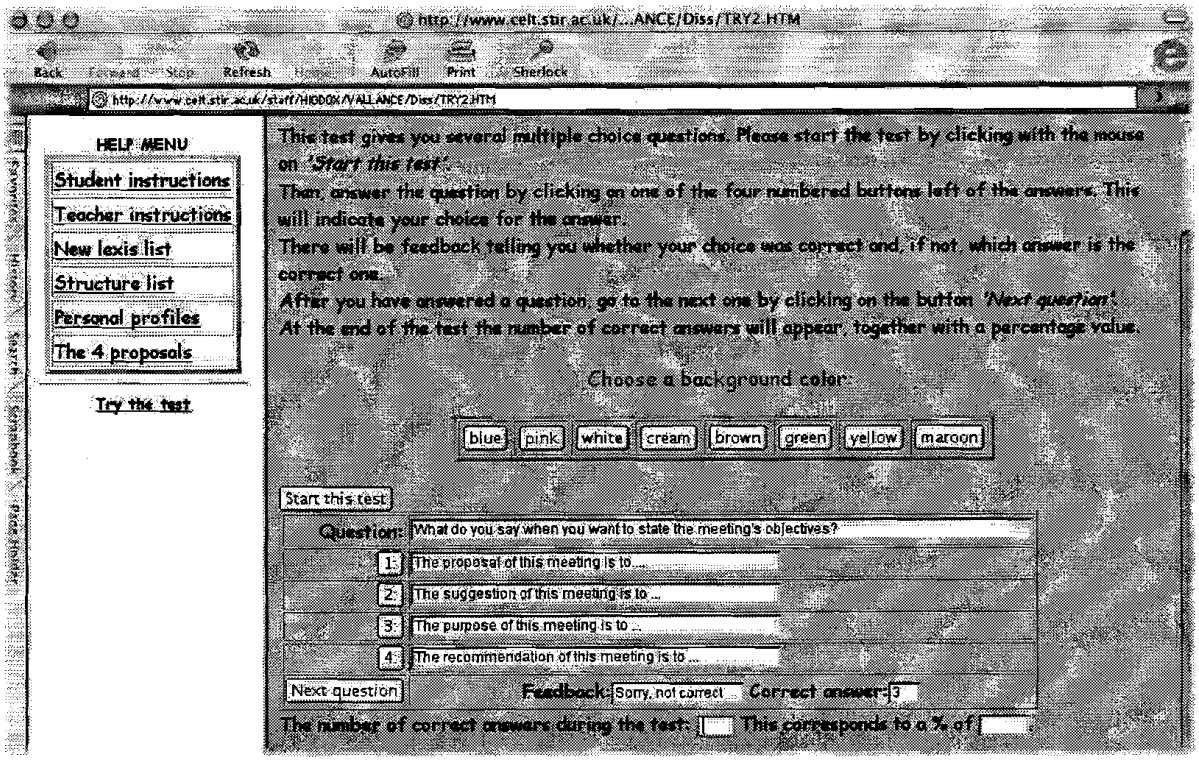

Fig. 4. An online multiple-choice test at the end of a business maze.

are collated server-side and securely stored for retrieval by the testing authority or teacher. World Class Arena is online at http://www.worldclassarena.org/.

Discussion

The aforementioned test types, so typical of many so-called e-learning programs are often associated with learning methods from the Behaviourist era (drill until mastery), hence aligned with traditional teaching methods. Also, the technical limitations of test designers tend to result in the adoption of the MCQs, cloze and word-order tests of language items. However, some pre-planning by the English teacher can bring about a more constructivist language learning environment where students' summative performance can be assessed and discussed, not necessarily by the computer but by the student's peers, the teacher or an observer.

For example, given that a typical English language lesson follows the Presentation – Practice – Production (P–P–P) format, a carefully constructed Hot Potatoes exercise can be used to present grammar. If the lesson's objective is to introduce, say, the passive form (e.g. "The graph was plotted. The project has been completed"), it is anticipated the students will have some knowledge of past tense, present perfect and active sentence constructions. They will therefore use this knowledge to develop a hypothesis of the construction of the passive sentence and its usage in meaningful contexts. Sentence construction or cloze passage completion, utilising surrounding lexical items as clues, can easily be designed by the teacher using Hot

Potatoes. A TexToys passage completion exercise, where all the words are initially hidden, can be developed in the same manner. Working in pairs around a computer, the students progress through the teacher-constructed computer exercise and they are encouraged to develop a hypothesis about the grammar, its meaning and its function. Once completed the students then move away from the computers and share their hypothesis with the class. The teacher guides the students, eventually resulting in a group hypothesis that can be tested. This will become the practice stage, which can be undertaken in a number of ways in this low stakes testing environment. Examples can be a listening passage where students retrieve information and construct related sentences, a reading maze for students to seek relevant information, or students can design a Hot Potatoes exercise to test their peers. The grammar is not tested in isolation but embedded within carefully constructed contexts.

Finally, students need to utilise the grammatical structure in authentic situations – the production or performance stage of the lesson. At this juncture a reading maze can be utilised where students collaborate to develop an engineering report. Students use the lesson's sentences and context to write the report. Choosing any of the aforementioned tools, the teacher may later test the students' learning by requesting the students to complete one final computer exercise. Scores can be tagged and utilised as a springboard for the next English lesson. No paper-based worksheets need to be physically graded by the teacher as the computer will provide an immediate and accurate score.

This lesson therefore encourages an inductive approach to learning where students collaborate, construct knowledge, form and test hypotheses, and meaningfully practice the target language. The computer tools are utilised to accommodate the process of learning and then later the product of the lesson by testing the students' perceived acquisition. Could this lesson be undertaken without a computer? Certainly, but it is ascertained that the tools utilised in this low stakes testing classroom adds value to the students' learning and the teacher's delivery of the lesson's objective.

Conclusion

The article has summarised computer use in high and low stakes testing in English language learning. High stakes testing was exemplified by CATs, which adapt to the test takers' input. It was explained that such tests are currently being utilised by commercial organisations but require skills far beyond the scope of English language teachers. More pragmatic usage of computers by teachers is in a low stakes testing environment. A number of personally designed examples were exemplified. Although such tests may be considered behaviourist, a lesson example was provided to illustrate how such tools can be utilised in a student-centred, constructivist learning classroom. The computers can be seen to add value to the lesson.

Roever (2001) in a recent article entitled *Web-based Language Testing* argues that computers cannot simulate communicative situations to test communicative competency, accuracy or performance. However, this will soon change. Users have been communicating with their computers for many years, from the ELIZA project in the 1960s to today's web-based Intelligent Agents or chat bots. With the technical developments of speech recognition, the future computer based test may indeed test a student's language appropriacy and accuracy over a network. Currently, IBM, Microsoft, Apple and Sun conduct high and low stakes testing and certification through network technologies. Given such utilisation and the continuing development of Intelligent Agents, it is predicted that CBT will become a major medium of test delivery to the more digitally literate population of the twenty-first century. Chat bots are online at http://www.botspot.com/bot/what_is_a_bot.html.

Michael Vallance has been a Lecturer at NIE since November 2000. He teaches "Using IT in the Language Classroom" to teacher trainees and also "Introduction to Computer Skills" to Chinese scholars. Prior to joining NIE he taught at Temasek Polytechnic for three years and spent seven years in Japan at Temple University and the International Language Centre. He has a Masters Degree in Computer Aided Language Learning and is currently pursuing a Doctorate in Education. His research interests include Educational Technology. He spends his limited free time managing a Singapore computing society. E-mail: mval@nie.edu.sg

References

Amrein, A. L. & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives*, **10**(18) [online]. Available: http://epaa.asu.edu/epaa/v10n18/.

Higgins, J. & Johns, T. (1984). *Computers in Language Learning*. London: Collins.

Madsen, H. S. (1991). Computer adaptive testing of listening and reading comprehension. In Dunkel, P. (Ed.), *Computer Assisted Language Learning and Testing: Research Issues and Practice*. USA: Newbury House.

Roever, C. (2001). Web-based language testing. *Language, Learning & Technology*, **5**(2), 84–94 [online]. Available: http://llt.msu.edu/vol5num2/roever/default.html.

Sackett, P. R., Schmitt, N., Ellingson, J. E. & Kabin, M. M. (2001). High stakes testing in employment, credentialing and higher education. *American Psychologist*, **56**(4), 302–318.

Stevenson, J. & Gross, S. (1991). Use of computerised adaptive testing model for ESOL/Bilingual entry/exit decision making. In Dunkel, P. (Ed.), *Computer Assisted Language Learning and Testing: Research Issues and Practice*. USA: Newbury House.

Warschauer, M. (1996). Computer assisted language learning: an introduction. In Fotos, S. (Ed.), *Multimedia Language Teaching*. Tokyo: Logos International, 3–20.

Windeatt, S. (1986). Observing CALL in action. In Leech, G. & Candlin, C. N. (Eds.), *Computers in English Language Teaching and Research*. London: Longman.