ReCALL **20**(1): 21-34. 2008 © European Association for Computer Assisted Language Learning
DOI: 10.1017/S0958344008000311 Printed in the United Kingdom

Language testing and technology: problems of transition to a new era

PATRICIA DOOEY
Department of Languages and Intercultural Education
Curtin University of Technology, GPO Box U1987, Perth, Western Australia 6845
(*email: P.Dooey@curtin.edu.au*)

Abstract

Technological advances have revolutionised methods of both teaching and testing in languages, and practitioners have eagerly embraced the opportunity to provide more innovative ways of doing this. The unique features offered by technology make it increasingly possible to test for a wide range of language skills required for a specific purpose. With the increasing need to test for English language proficiency and the importance placed on this facility, technology is being utilised to address issues of practicality, speed and efficiency. However, such advantages should not be embraced without due consideration for the essential qualities of any test; validity and reliability. With the inevitable shift towards computer-based testing, certain areas need special consideration. While computer-based tests can provide constructive diagnostic information to complement the language learning process, they should be used more selectively in other contexts, for example in high-stakes tests, examples of which are International English Language Testing System (IELTS) and Test of English as a Foreign Language (TOEFL). These tests provide an assessment of English language proficiency, a key component of university admissions criteria. This paper examines a number of issues related to the design and application of computer-based tests, with particular reference to construct validity, computer familiarity and practicality. It is recommended that in the short term at least, test-takers be offered the choice of test medium in the interests of fairness and equity.

Keywords: Computer-based language testing, Assessment, English, language proficiency.

1 Introduction

Testing has been described as "a universal feature of social life" (McNamara, 2000:3). All tests are important because they provide information about the test-takers, which may be used for a range of purposes. Because of the important uses of test results, there is a certain element of power associated with those who design and administer tests, contrasted with the powerlessness and anxiety of test takers. It is therefore crucial that test designers produce fair and accurate tests.

https://doi.org/10.1017/S0958344008000311 Published online by Cambridge University Press

"As one type of measurement, a test necessarily quantifies characteristics of individuals according to explicit procedures. What distinguishes a test from other types of measurement is that it is designed to obtain a specific sample of behaviour" (Bachman, 1990:21–2). When looking at the role of testing in any situation, the primary question to ask is 'for what purpose is the test being used?' The use of language tests has always been important because the results obtained may serve to enhance and promote the learning process. In the case of English language tests in particular, there is increasing pressure to achieve desired results, because such success provides gateways of opportunity for ever-growing numbers of learners.

> For millions around the world, English language competence is the key to information, educational opportunity and employment. In ESL testing our purpose should be to help people realize their educational and career goals, while assisting institutions in making the resource allocation decisions they must.(Braun, 2000:271).

In the case of overseas non-native speaking students wishing to study in an English-medium university, a range of tests is used to assess their language proficiency, so that decisions can be made about their future study. The extent to which the outcomes of a test can affect the candidates' futures is referred to as the stakes involved.

> The term high stakes is generally used to refer to a test upon which candidates' career or study plans hinge. Examples of such tests are TOEFL and IELTS when used as selection tests for tertiary study, or a language test required of migrants in order to pursue their profession in the new country. A low stakes test is one where the outcomes are less likely to affect the candidates' future lives substantially, for example placement or classroom achievement tests (University of Cambridge Local Examinations Syndicate, 1999:185).

Both the TOEFL and the IELTS have been widely used for a number of years to screen for English language proficiency; the former was established primarily for students applying to North American universities, while the latter was established in late 1989 as a joint venture by a British/Australian team, to replace its predecessor, the ELTS test. IELTS currently operates in over 400 test centres in 120 countries, (*IELTS Test Centres and Examiners*, 2006) and is used for admission to English-medium universities all over the world.

As developments in information and communication technology (ICT) progress rapidly, computers are playing an increasing role in language teaching and testing. With an increasing demand worldwide to learn English, it is not surprising then, that language test developers are taking advantage of the available technology to capture, store, process and retrieve the enormous amount of information needed to efficiently run a range of different language tests. Therefore technology can be incorporated into many types of language tests because of the flexibility and efficiency it offers, depending on how the administrators and designers choose to utilise it.

Language test developers must always be mindful of how they design and use tests. However, incorporating the additional dimension of technology offers special challenges in this regard because of the possible paradigm shift in our approach to language testing, prompted by technological change (Jones, 2003:3). This paper will

https://doi.org/10.1017/S0958344008000311 Published online by Cambridge University Press

look at the role of technology in language testing and will attempt to answer the following questions:

- In what ways can computerized tests improve traditional test methods?
- To what extent are the abilities tested on computer-based tests comparable to their pen-and-paper counterparts?

Reference will be made to a selection of specific studies which highlight the need for special considerations to be taken into account when moving towards computer-based testing (CBT), particularly in the case of high-stakes tests.

2 Language test characteristics

Whatever the purpose or method of any given test, there are two essential qualities required – it must be valid and it must be reliable. A test is valid if it measures what it is intended to measure. "The validity of a language test is therefore established by the extent to which it succeeds in providing an accurate concrete representation of an abstract concept (for example, proficiency, achievement, aptitude)" (University of Cambridge Local Examinations Syndicate, 1999:221). Closely connected with this is the concept of test reliability, which has been defined as "the actual level of agreement between the results of one test with itself or with another test" (ibid, 1999:168).

It is obvious then, that the more important the consequences of the test are, the more important it is to ensure validity and reliability. This is crucial in the case of high stakes tests, where serious decisions are being made on the basis of results obtained. In addition to the two qualities mentioned above, a test must be practical; practicality refers to a range of basic issues, including having the infrastructure and resources readily available to administer the test efficiently and economically.

Technology is increasingly being promoted as a powerful force which is rapidly changing the face of education, and since language testing is an integral part of language learning, it is not surprising that language testing research is turning more and more towards computer-based testing (CBT). Although this concept has existed since the mid-1980s, it is only relatively recently that computer-based testing has been adopted by the governing bodies of the larger scale tests. For example, in 1998 the Educational Testing Service (ETS) introduced the computer-based TOEFL to enhance assessments by using electronic technology to test more complex skills (ETS, 2003:3).

Tests that are administered via a computer terminal are known as Computer Assisted Tests (CATs). These are relatively easy to run, and involve receptive-response items, including multiple-choice items or true/false items. They are therefore easy to adapt to the computer medium. A specific type of computer assisted language test is the Computer-Adaptive Language Test (CALT). The CALT was developed using a general measurement theory known as Item Response Theory (IRT). IRT is based on a probabilistic theory, that is, it calculates the probability of a person getting a particular item right (Alderson, Clapham & Wall, 1995). Extensive theoretical research was carried out on CAT and IRT in the 1960s and 1970s by the U.S. Department of Defence, which recognised the potential benefits of adaptive testing (Dunkel, 1999). In CATs, the candidate's ability is monitored as s/he goes through the test, and suitable items are

https://doi.org/10.1017/S0958344008000311 Published online by Cambridge University Press

selected which most match his/her ability. The items are stored in an 'item bank' and the computer selects appropriate items as the test proceeds. Once the candidate reaches a particular level (which is determined according to a specified standard error of measurement), the test is automatically terminated. The system is efficient in that it provides an instant result, using the minimum of items necessary, and in a shorter space of time than with its pen-and-paper counterpart. Its precision lies in the degree of accuracy with which the candidates' level can be estimated. The advantages of this system are obvious – the flexibility of administration, the shorter period of time needed to take the test, the availability of diagnostic information to both teachers and students, and finally, a more efficient use of test bank items (because of the fact that exposure is limited to only those items necessary to determine a particular candidate's level). By the 1980s, CATs were being used by many language institutes and at universities in many parts of the world, and modern developments in technology and computer software have made computer-adaptive testing become a viable alternative to conventional pen-and-paper testing. In the language testing situation, CALTs (computer adaptive language tests) offer the same advantages, as noted by Madsen (1991:237).

Technology promises a range of capabilities, including speed, accuracy and efficiency – in the case of language testing, technology is being used partly because it is available, and partly because it is intended to offer improvements in the way the job has been done traditionally. However, while there is little doubt that computer-based testing has in many ways rendered the language testing process more practical, the move to embrace technology needs to take place whilst at the same time conserving the underlying basic qualities of validity and reliability in language testing. For example, Green and Maycock (2004:4) draw our attention to one prevalent view "that technology has a narrowing, constraining influence: that it imposes certain solutions on language testing simply because they *are* possible, rather than valid".

There is some concern therefore, that the validity of computerized language tests may be compromised, as the second language (L2) field has long promoted performance-based assessment, a form of assessment that does not lend itself as easily to computer administration as do more traditional test formats (Chalhoub-Deville, 2001). For example, computers may not be able to offer open-ended communication, which demonstrates characteristics which are as closely related to real-life as possible, such as in the case of a one-on-one interview. Thus it is important to ensure that the test design continues to incorporate the same generic principles underpinning test construction, notwithstanding the extent to which technology is used. However, the increasing use of technology in language tests has raised a number of issues worthy of consideration, including construct validity, computer familiarity and practicality.

3 Language testing and technology – the implications

In this section, a number of issues related to using technology in language testing will be identified and discussed. For each one, a definition will be provided, together with an explanation of its application in the context of ICT. Finally, reference will be made to the findings of a number of relevant research studies, which report the positive and negative outcomes of computerized testing, and highlight the specific challenges involved for test users; that is, test takers, test designers and test administrators.

https://doi.org/10.1017/S0958344008000311 Published online by Cambridge University Press

3.1 Construct validity

Firstly, of primary concern to computer-based tests is the issue of construct validity, and in particular the extent to which construct validity is maintained with a computerized test. "Construct validation involves an investigation of the qualities that a test measures, thus providing a basis for the rationale of the test" (University of Cambridge Local Examinations Syndicate, 1999:33). In the case of computer-based language testing, it is important that the test is measuring the pertinent language skill or ability of the test-taker, rather than his/her computer skills. In particular, the mode of presentation should not have an effect on the test-taker's results; for example, reading from a computer screen is different from reading from a double-page spread in print, and requires different cognitive skills. Furthermore, some test-takers "may experience fatigue when reading extended passages on computer" (Blackhurst, 2005:16). Another point to note is that there is a fundamental difference between handwriting and word-processing. Therefore, the equivalence of corresponding conventional and computerized test forms must be established, thus taking not only the validity, but also the reliability of the test, to task. The important question here is: can the same skills be equally and as fairly measured by computer tests, as by their pen-and-paper counterparts? While the score equivalence of computer-based and paper-based tests has been the subject of recent research, "individual differences have the potential to significantly affect the equivalence of computer- and paper-based assessments" (McDonald, 2002:309).

In the context of large-scale high-stakes tests which are currently moving towards becoming computer-based, this is a key issue. A prime example of this is TOEFL. The TOEFL test is a standardised test of English proficiency. It was developed in 1964 and is administered by the ETS in the USA. It is most commonly used by non-native speakers of English who wish to study at North American universities, but is recognised and administered world-wide. Major changes were undertaken in the implementation of TOEFL in the late 1990s. The most significant was the changeover from the traditional pen-and-paper test to a wholly computer-based format, where the reading and structure sections are computer-adaptive. The test designers needed to ensure, therefore, that test takers would not be disadvantaged by lack of computer familiarity.

As part of this change, a major quantitative study was undertaken by Taylor, Kirsch, Eignor and Jamieson (1999) to look at the relationship between computer familiarity and performance on the computer-based test. The prime reason for this was to eliminate any possibility that language proficiency would be confounded with computer proficiency, as this would undermine the construct validity of the test. The subjects were all TOEFL test takers from twelve international sites. A group of 1,200 examinees was categorized as 'low computer-familiar' and 'high computer-familiar'; this was based on access, (or where the TOEFL examinees use computers), self-assessment of attitude and ability, use of and experience with computers, and use of and experience with related technology. They then worked through a computer tutorial and a set of sixty computerized test items. For purposes of ensuring that TOEFL test takers were sufficiently familiar with computers to take the test, the test developers (ETS) designed a tutorial to be taken by all test takers. The authors concluded that "no evidence exists of an adverse relationship between computer familiarity and computer-based TOEFL test performance due to lack of prior computer experience" (Taylor *et al*., 1999:219–20). However, they did raise concerns about future generations of computer-based tests: "For

https://doi.org/10.1017/S0958344008000311 Published online by Cambridge University Press

example, at present, little is known about the effects of various multimedia applications or how what is being measured actually changes when presented in a CBT environment rather than a paper-and-pencil environment" (Taylor *et al*., 1999:270). Finally they recommended further research into the area of whether what is being measured actually changes when presented in a CBT environment, as test items and the level of manipulation become more sophisticated over time.

Using the Test of English Proficiency developed by Seoul National University (TEPS) as a model, another investigation was carried out into the comparability of a paper-based language test and its computer-based equivalent. The findings supported comparability in listening comprehension, grammar, vocabulary and reading comprehension; however the reading showed the weakest comparability (Choi, Kim & Boo, 2003).

In a review of the literature addressing the comparability of pen-and-paper tests and computerized tests of reading, and to identify pertinent areas for future research, Sawaki (2001) raised some important issues regarding the mode of presentation in language assessment. While one section of the review focused on first language (L1 reading research, rather than second language (L2), the author started out with the same basic premise; that reading from a computer screen is becoming more commonplace in our daily lives, and raised the question of whether reading from a computer screen is the same as reading in print, specifically in the case of L2 readers. Sawaki believed that the current push towards computerized testing is moving ahead without sufficient empirical evidence in this area, and raised concern over whether the same construct is being measured in the computerized and equivalent pen-and-paper tests. The author therefore looked at:

- Studies that address general construct validity issues of computerized tests in cognitive ability as well as language assessment.
- Studies that shed light on the effect of mode of presentation on reading performance conducted mainly in ergonomics, education, psychology, and L1 reading research (Sawaki, 2001:38).

In a fairly extensive review of the literature relating to the first point above, the author selected critical areas for comparison between computerized and pen-and-paper tests. These included content, psychometric equivalence, the impact of computerisation on examinees and finally the decisions to be made about examinees on the basis of results. However, the author found that the assessment literature in this area indicates that "the empirical findings as to comparability of conventional and computerized tests are rather mixed" (Sawaki, 2001:6). Whilst there were some differences, they did not seem to be of any great consequence in terms of the decisions being made.

In relation to the studies reported in the second point above, the conclusions were once again mixed. While it was found that comprehension of computer-presented texts is generally as good as that of printed texts, it also seemed that reading speed may or may not be affected by mode of presentation. Despite the mixed findings reported in the review, the author clearly identified some crucial areas worthy of further investigation in her discussion. These include the notion of addressing the effect of mode of presentation on the process rather than the product of reading, and to this end, suggested methodologies such as analysis of eye movement and verbal protocol analysis.

https://doi.org/10.1017/S0958344008000311 Published online by Cambridge University Press

Sawaki highlighted the dearth of interpretive research, for example post-hoc interviews and questionnaires in the studies reviewed. She also identified two important limitations in the scope of the literature. These are: firstly, the fact that the literature did not investigate issues that are more important to longer texts which would be used in more advanced L2 reading tests (most of the studies reviewed involved shorter texts or proof-reading); secondly, Sawaki pointed out that the study did not cover empirical studies on reading passages that were accompanied by visual prompts, such as figures or graphics. She commented that "if such visual prompts are incorporated into a computerized test, effects of these visuals on examinees' cognitive processing and the potential effects of performance differences across modes should also be addressed" (2001:13).

Sawaki concluded that the effects of mode of presentation on test performance are indeed complex, and that the literature suggests that these effects:

> may be observed in a change in the nature of a test task, in a decision based on a test score, in test completion time, and in test takers' affect, for example, whereas the test score itself may not necessarily be influenced…further investigation in this area should be an integral part of future construct validation in computerized tests of reading in a second/foreign language (2001:13).

These conclusions have important implications for those who use computerized language tests. Firstly, even where the final score may not be affected, test-takers may be employing different strategies and skills in order to reach the same end. Therefore the skills that are being demonstrated must match the skills that are being thus tested in order for the test to maintain validity. Secondly, the management of time plays a crucial role in language testing, particularly in high stakes tests; therefore comparative studies of time taken to complete computerized versus pen-and-paper based tests would help to ensure parity between the two formats, given that time management is crucial in high-stakes tests.

3.2 Computer familiarity – the broader issues

The study undertaken by Taylor *et al*. (1999), as reported earlier, highlighted the need to be computer familiar, in the narrow sense, in relation to computer-based tests. However, the broader interpretation of computer familiarity, involving issues of attitudes towards computer use, levels of anxiety which could be associated with computer use, and issues of access and equity was also identified as needing further research.

One important issue for language test designers planning to use computers in testing is that of attitudes in general to computer use. Anxiety in testing situations can be compounded by (perceived) lack of expertise when technology is being used. As part of the preparation for the computer-based TOEFL, a study was undertaken by Jamieson, Taylor, Kirsch and Eignor (1998), which involved the construction of a set of tutorials designed to prepare examinees, regardless of computer familiarity, to take the computerized version of TOEFL, which was scheduled to go 'live' in July 1998. These tutorials were trialled as part of the computer familiarity study reported earlier.

Information gathered from 1,169 of the examinees who participated in that study was categorised and analysed according to timing, performance data and self-reported

https://doi.org/10.1017/S0958344008000311 Published online by Cambridge University Press

attitudes. Analyses took into account both English language ability and computer familiarity. In all cases, even with those who were computer-familiar, the tutorial took longer than anticipated to complete. In terms of performance, over 95% of the examinees completed the exercise successfully; little or no remediation was needed by either group, (i.e. computer familiar or computer unfamiliar) and remediation, when used, was 100% successful. However, the third of these areas was the most interesting. Attitude towards ease of use of computers was compared with information taken from questionnaires which had been administered previously, and it was found that there was a positive shift in this respect. In other words, examinees appeared to become more comfortable with using the computerized TOEFL test. However, a large proportion of the overall examinee population still reported some unease. Furthermore, English ability did not seem to have a major effect on the change of comfort level. Although it is not clear whether the respondents were anxious about taking the test itself, or taking a computer-based version, or about using computers *per se*, this area is worthy of further investigation, given that TOEFL is such a high-stakes test. There is little doubt that anxiety can play a significant part in test performance, particularly anxiety related to the need to manipulate new technology. This was also noted by Madsen (1991:251) who reported the results of an affect questionnaire administered at Brigham Young University in conjunction with a computer-adaptive test of listening and reading, finding that "In short, experience with computers was not found to alleviate test anxiety".

Although it is generally agreed that issues of computer familiarity are becoming less important now than in the past (Sawaki, 2001; Taylor, Jamieson & Eignor, 2000), what has emerged is that, in general, access to computers is not equally divided among different nationality groups. This means that while access to computers worldwide is increasing, this trend is not as rapid in certain parts of the world. Furthermore, certain language groups seem to be less familiar with computers than others. This highlights the widening gap between the computer 'haves and have nots'. Therefore care must be taken not to expose some groups of students to a potential bias against those who are not yet fully comfortable with computer technology. Such a bias, based on computer familiarity, could have serious implications for the future validity of TOEFL, given that it was generally presented as a computer-based test in 1998, with no option for candidates to choose the traditional pen-and-paper version (with the exception of the writing component of computer-based TOEFL, which allowed the choice of handwritten or word-processed responses). It also raises questions of equity, in an environment where increasing numbers of African students, for example, are taking up the opportunity to study abroad, and often need to take or repeat the TOEFL test in the host country. Although the researchers reported these issues as limitations, lack of computer familiarity was predicted to diminish over time. Consequently, when the TOEFL test introduced the changeover from computer-based to iBT (internet-based) in 2005, the requirement to type became compulsory, since the "vast majority of test-takers on the computer-based TOEFL test choose to type their answers" (Educational Testing Service, 2005: 23). Nonetheless, in the immediate future at least, candidates should not be disadvantaged by not having a choice in the test medium.

With more and more non-native English speaking students wishing to study overseas, Taylor *et al*. (2000) decided to investigate the issue of whether the increasing use of

https://doi.org/10.1017/S0958344008000311 Published online by Cambridge University Press

computers in higher education as well as in language instruction, could cause problems for those students who did not have a background of computer use:

> The increasing use of computers in North American higher education as well as in language instruction and assessment raises concerns about access and equity for many international students, who may be caught on the wrong side of the digital divide (Taylor *et al*., 2000:575).

The researchers were interested in the level of computer familiarity among those students who were taking the TOEFL test, for admission to North American universities, as a representative sample of English language learners around the world. They were also interested in finding out whether or not the profile (that is, of computer familiarity among students hoping to participate in higher education in North America) was changing. An extensive literature review revealed which countries could be classified as the 'haves' and the 'have nots' in this respect, and much of the literature focussed on the use of computers in schools (from elementary through to upper secondary). However, this did not address the area of particular interest – the extent to which English language students wishing to study in North America had previously accessed computers on a regular basis. While local students were deemed to be familiar with computers and to use the Internet regularly, concern was raised about the incoming overseas student population. A range of questions had to be addressed, including:

> Are all international students equally familiar with computers? Do these students know how to use an English word-processing application? Do all of these students arrive on campus already knowing how to use the Internet? In addition to English proficiency, international students will increasingly need computer skills to be ready for college. Are such expectations realistic for all matriculating international students (Taylor *et al*., 2000:577)?

In an attempt to address the issue of computer familiarity, a questionnaire was administered twice to all TOEFL examinees, in April/May 1996, and again about twenty months later in 1997. Then it was returned to the ETS for analysis. Responses were compared across test centre region, and by native language; a substantial proportion of the test-takers came from Asia, representing seven major language groups.

In each of the three areas of interest, (namely, frequency of computer use, frequency of use of word processors, and frequency of use of the Internet), increases in usage were noted in the 20-month interval, especially in the use of the Internet. This was good news for those developing computer-based language teaching and testing materials. However, the authors were frank about the limitations of the study. Firstly, it was designed to be descriptive and to show general trends, therefore the data should be interpreted broadly. In particular, the results should not be considered indicative of a high level of computer familiarity among TOEFL test-takers, because access to technology is not evenly distributed across sub-groups. For example, while a marked increase appeared overall in computer use over the 20-month interval between the two survey administrations, the use and increase were considerably less for African students. The researchers concluded that "one quarter to one half of the students from most regions of the world will likely

https://doi.org/10.1017/S0958344008000311 Published online by Cambridge University Press

need help learning how to use English word-processing programs and the Internet once they arrive at North American colleges and universities" (Taylor *et al*., 2000:584). Although they acknowledged that the data provided by the survey were already dated, they recommended that teachers in EAP (English for Academic Purposes) programs should continue to "assess, rather than assume, the computer familiarity of their students" (Taylor *et al*., 2000:584).

Finally, Maycock and Green (2005) reported on the results of a questionnaire administered in conjunction with a Computer-Based IELTS (CB IELTS) trial, in which candidates' attitudes towards CB IELTS were sought, particularly in relation to a comparison of their performance on Paper-Based (PB) and CB IELTS. When asked about whether or not they believed computer familiarity would be an advantage, "67.7% agreed that candidates with good computer skills would perform better on CB IELTS than those with only basic skills" (op. cit.: 4). With this in mind, it would seem to be more equitable to leave the choice of medium to the individual.

For institutions wishing to accept international students, there are two issues at stake here. Firstly, where computers are being used for assessment purposes, all students must be on an equal footing with regard to the ease with which they can use computers, therefore they need to be comfortable with the medium. Secondly, it must be established that such students are in an environment where they are using computers on an ongoing basis, either for reference purposes or for completing assignment work. In order to do this, universities may need to not only provide ongoing support in this regard, but also to monitor its use, so that international students are not disadvantaged because of lack of previous experience.

3.3 Practicality

When using technology for any language test, the question of practicality must always be considered, particularly in the case of large-scale high stakes tests. In the first place, the test must be viable and cost-efficient, and the equipment needed to run the test efficiently must be readily available. In other words, sufficient up-to-date computer terminals should be provided, and be in working order. When the computer-based format for TOEFL was first introduced in 1998, many well-established testing centres in the US and around the world were no longer able to run the test because of the lack of the necessasry infrastructure. Even with the necessary hardware, the cost of running such a large-scale test could be prohibitive for some centres. On an even more pragmatic level, the consequences of the inevitable unforseen temporary failure of technology on a high-stakes test could not only be serious, but costly. Even in the most technologically advanced institutions, the possibility of technical malfunction, however remote, is always a possibility that needs contingency plans put in place.

On the other hand, computer assisted tests have proven to be immensely more practical in many ways than their pen-and-paper counterparts. The overwhelming benefit of the computer-adaptive test, for example, is the flexibility with which it can be administered, on a walk-in basis, for even a small group of students, with no need to host a single large administration at a pre-arranged time. There is also the added advantage of the reduction in time taken to complete the test, a unique feature of the computer-adaptive test.

https://doi.org/10.1017/S0958344008000311 Published online by Cambridge University Press

In terms of item-banking, the computer is an ideal medium for storage of large banks of test items ready for instant retrieval, depending on the ability of the individual examinee. On the one hand, the way such a test is set up means that each test is unique, since no two tests will present exactly the same set of items. On the other hand, however, because of the frequent exposure to such items due to small, *ad hoc* administrations (as opposed to large, fixed time administrations), it is necessary to keep a broad bank of items available at any given time. Furthermore, on some computer-based tests omissions are allowed, while on others candidates are not permitted to proceed to the next item until the previous one is answered. If omissions are allowed, this could lead to problems resulting from exposure to a greater range of items. This could affect the integrity of the test.

Computer-adaptive tests can also offer diagnostic feedback to students in situations where that is the purpose of the tests. This feedback can pinpoint the ability of the student with great precision through IRT. Because each test is unique, security is enhanced, with the elimination of possible repeated exposure to items. CATs are also designed to allow the students to work at their own pace and their own level, thus causing less frustration and anxiety. In this way, such tests are more learner-centred, because of the fact that they can be uniquely tailored to the individual's level, rather than being too easy or too difficult, as might be the case with fixed format tests. As such, they can be used to combine language teaching and testing, providing immediate feedback on progress made. This is particularly useful for low-stakes tests.

Another practical application of CATs is the speed and accuracy of computerized scoring for clerically marked test components, where human raters can be substituted by technology. The advantages of this system are that the possibility of human error is eliminated, and it is much cheaper.

Finally, computer-based assessment offers the capacity to store enormous amounts of data which can be used later for research purposes: for example, item analysis, details about the time taken to answer questions and levels of difficulty of test items. Overall, therefore, CATs offer substantial enhancement to the language testing process when used appropriately. All in all, there is no single best test, but there is a most appropriate test for every purpose, and this may comprise a combination of test methods.

4 The impact of technology in testing productive and receptive skills in high stakes tests

With the recent launch of Computer Based IELTS (CB IELTS) in certain centres, a number of trials have been undertaken by the Research and Validation group at Cambridge ESOL, one of the IELTS partners. The purpose of these trials was to establish whether or not CB IELTS is comparable to its pen-and-paper counterpart (PB IELTS). In relation to the receptive skills of Listening and Reading, "The evidence gathered to date suggests that CB IELTS can be used interchangeably with PB IELTS and that candidates, given adequate computer familiarity, will perform equally well on either version of the test" (Green & Maycock, 2004:6). A later study by Maycock and Green (2005) supported the idea of comparability between computer-based and pen-and-paper based versions of the IELTS test, citing O'Sullivan, Weir and Yan (2004: 9) who cautioned that "although the evidence points to the view that computer familiarity

https://doi.org/10.1017/S0958344008000311 Published online by Cambridge University Press

alone may not have a significant effect on CB test results ... It cannot be ignored when comparing [PB] and [CB] tests". Once again, there is evidence to suggest that this comparability is contingent upon a certain level of computer familiarity.

In relation to marking the writing component of IELTS (which can be typed or handwritten), issues of scoring and the treatment of errors would need to be dealt with through examiner training and guidance (Green & Maycock, 2004:6). The speaking component of IELTS will continue to be run via face-to-face interviews. Therefore, the role of specialist trained examiners in the testing process remains critical in high stakes tests.

In the direct writing component of the TOEFL test, Wolfe and Manolo (2004:53) undertook a study which "sought to determine whether performance on a direct writing assessment is comparable for examinees when given the choice to compose essays in handwriting versus work processing". The study was based on the computer-based TOEFL, which offered candidates the choice of medium, handwritten or word-processed, in the writing section (this format has since been replaced by iBT, which requires the candidates to type only).

The researchers were particularly interested in the effects of using such a test on a diverse population of international test-takers; for example, those from developing regions, those who speak languages with characters that are different from the Roman or Cyrillic alphabet or those with lower levels of English proficiency. They looked at the possibility that some candidates may be exposed to what the researchers termed the 'double translation' effect; in other words, they might have to translate from their own language into English and then from English into the appropriate keystrokes. Their data "suggest that demographic characteristics that one would suspect to be related to computer exposure are indeed related to performance on a computer-based direct writing assessment, albeit only weakly" (Wolfe & Manolo, 2004:61), and concluded that "those involved in developing, administering, and reporting results of direct writing assessments designed for examinees with diverse language background should think seriously about providing examinees with a choice of composition medium" (*op. cit.*: 21).

5 Conclusion

Returning to the fundamental questions posed at the beginning of this essay, it would seem that CATs may well improve on pen-and-paper versions, but only in certain circumstances. From a pedagogical perspective, individual students can receive detailed diagnostic feedback on their performance, and greater accuracy can be achieved on scoring selected-response items. Limited exposure to masses of test items is made possible through item-banking, which has the dual advantage of preserving test integrity, while offering candidates a test made up only of those items with the appropriate levels of difficulty, as in the case of adaptive tests. From a practical perspective, CATs take less time, and candidates can work at their own pace, thus experiencing less frustration. Tests can also be taken on demand, thus eliminating the need to set up large-scale test administrations.

Technological advances have offered improvements in all walks of life, and language testing is no different; however, measuring these benefits is not so straightforward. In the case of high-stakes tests such as IELTS and TOEFL, research so far shows that in terms of test performance there is an acceptable degree of comparison between scores

https://doi.org/10.1017/S0958344008000311 Published online by Cambridge University Press

obtained on computer-based tests and their pen-and-paper counterparts. However, this situation assumes a certain level of computer familiarity, and it is evident that the 'digital divide' still separates many of our international students.

Furthermore, there is some evidence to suggest that different abilities are being measured across the two media, even where the scores obtained could be considered comparable. So the area of construct validity needs further research in terms of the skills being measured. On any test, the scores obtained should represent a realistic estimate of the skills and abilities being tested. The degree to which CATs can achieve this in the productive skills of writing and speaking needs further research. In particular, communicative language competence will always require the intervention of human scorers, or raters for marking. Furthermore, theoretical studies of writing suggest that there are fundamental differences between writing by hand and word-processing. However, what is important to ascertain here is whether or not this would have an effect on the marks obtained.

While technology is already revolutionising the world of education and its related fields, there is always the danger that it will fail. For this reason, using a single medium of testing which relies heavily on technology can be both costly and risky. As a consequence, the pen-and-paper test versions should always be an option, particularly in high-stakes testing. This not only provides alternatives for those administering the test, it also offers test-takers the choice in how they take the test. For example, in addition to the computer-based version, all IELTS centres "will continue to offer pen-and-paper based IELTS and candidates will be given the choice of the medium in which they wish to take the test" (University of Cambridge ESOL Examinations, 2006:3). This is important for those who wish to study in English-medium universities, but who have had limited access to computers.

Finally, the issue of access and equity needs to be considered. Although the literature indicates an increasing level of computer use and familiarity in most of the overseas student population, research has cautioned that access and use is not evenly distributed across different groups of students. For example, it is evident that a certain proportion of African students in particular come from an environment which lacks the infrastructure needed to allow them regular access to computers. This has implications not just at the testing level, but throughout those students' academic careers. Since this sector of the overseas student population in English-speaking countries is growing, provision will need to be made for them to receive the necessary instruction, if they are to stand on equal ground with students from other countries.

Ultimately, the ways in which tests are used should be determined by the test-users themselves, bearing in mind the purpose for which the test is being used, the range of skills it is intended to test, and the best and fairest means of testing those skills. "Test-users need to be cognizant of the properties of the instruments they employ and ensure appropriate interpretation and use of test scores provided" (Chalhoub-Deville & Turner, 2000:537).

References

Alderson J., Clapham, C. and Wall, D. (1995) *Language test construction and evaluation*. New York: Cambridge University Press.

https://doi.org/10.1017/S0958344008000311 Published online by Cambridge University Press

Bachman, L. (1990) *Fundamental considerations in language testing*. Oxford: Oxford University Press.
Blackhurst, A. (2005) Listening, Reading and Writing on computer-based and paper-based versions of IELTS. *University of Cambridge ESOL Research Notes*, **21**: 14–7.
Braun, H. (2000) A post-modern view of the problem of language assessment. In: Kunnan, A. J. (ed.) *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida*. Cambridge: Cambridge University Press, 263–72.
Chalhoub-Deville, M. (2001) Language Testing and Technology: past and future. *Language Learning & Technology*, **5** (2): 95–7.
Chalhoub-Deville, M. and Turner, C. (2000) What to look for in ESL admission tests: Cambridge certificate exams, IELTS, and TOEFL. *System*, **28** (4): 523–39.
Choi, I., Kim, K. and Boo, K. (2003) Comparability of a paper-based language test and a computer-based language test. *Language Testing*, **20** (3): 295–320.
Dunkel, P. (1999) *Considerations in Developing and Using Computer-adaptive Tests to Assess Second Language Proficiency*. Washington DC: ERIC Clearinghouse on Languages and Linguistics. (ERIC Document Reproduction Service No. ED435202)
Educational Testing Service (2003) *TOEFL Test and Score Data Summary*. Princeton, New Jersey:Author.
Educational Testing Service (2005) *TOEFL iBT Tips*. Princeton, New Jersey: Author.
Green, T. and Maycock, L. (2004) Computer-based IELTS and paper-based versions of IELTS. *University of Cambridge ESOL Research Notes*, **18**: 3–6.
IELTS – Test Centres and Examiners (2006)
http://www.ielts.org/testcentresandexaminers/article3.aspx
Jamieson, J., Taylor, C., Kirsch, I. and Eignor, D. (1998) Design and evaluation of a computer-based TOEFL tutorial. *System*, **26** (4): 485–513.
Jones, N. (2003) The Role of Technology in Language Testing. *University of Cambridge ESOL Research Notes*, **12**: 3–4.
Madsen, H. (1991) Computer-adaptive testing of listening and reading comprehension. In: Dunkel, P. (ed.) *Computer-assisted language learning and testing: Research Issues and Practice*. New York: Newbury House, 237–57.
Maycock, L. and Green, T. (2005) The effects on performance of computer familiarity and attitudes towards CB IELTS. *University of Cambridge ESOL Research Notes*, **20**: 3–8.
McDonald, A. (2002) The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers & Education*, **39**: 299–312.
McNamara, T. (2000) *Language Testing*. Oxford: Oxford University Press.
O'Sullivan, B., Weir, C. and Yan, J. (2004) Does the computer make a difference? unpublished IELTS Research Project Report.
Sawaki, Y. (2001) Comparability of Conventional and Computerized tests of Reading in a Second Language. *Language Learning and Technology*, **5** (2): 38–63.
Taylor, C., Jamieson, J. and Eignor, D. (2000) Trends in computer use among international students. *TESOL Quarterly*, **34** (3): 575–85.
Taylor, C., Kirsch, I., Eignor, D. and Jamieson, J. (1999) Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning*, **49** (2): 219–74.
University of Cambridge ESOL Examinations (2006) *IELTS Handbook 2006*. Cambridge: UCLES.
University of Cambridge Local Examinations Syndicate (1999) *Dictionary of language testing*. Cambridge: Cambridge University Press.
Wolfe, E. and Manalo, J. (2004) Composition medium comparability in a direct writing assessment of non-native English speakers. *Language Learning and Technology*, **8** (1): 53–65.

https://doi.org/10.1017/S0958344008000311 Published online by Cambridge University Press