6 Computers in language testing

Glenn Fulcher, *University of Surrey*

Glenn Fulcher has worked in TESOL as a teacher and teacher trainer overseas and in the United Kingdom since 1982. He gained his MA in Applied Linguistics from the University of Birmingham in 1987, and his PhD from the University of Lancaster in 1993. He has a special interest in Language Testing, and is currently on the TOEFL Committee of Examiners. He has also served on the Executive Board of the International Language Testing Association. He is currently Director of the English Language Institute at the University of Surrey, where he teaches language testing and Applied Linguistics on the MA in TESOL and MSc in English Language Teaching. g.fulcher@surrey.ac.uk

Introduction

Computers have played a key role in language testing since 1935. From the early scoring devices to the latest Computer Adaptive Tests (CATs), computers have come to play a major role in test construction, item banking, test administration, scoring, data analysis, report generating, research, and the dissemination of research. As the machines and software have become more sophisticated, a range of practical, research and ethical issues have arisen. This chapter will outline the role of the computer in language testing and discuss some of the complex issues that need to be addressed in the first decades of the 21st century.

The IBM model 805

The first recorded use of a 'computer' in language testing dates back to 1935 when the IBM model 805 became commercially available. It was the first machine capable of scoring objective tests, and was immediately put into use in the United States to reduce the labour intensive and costly business of scoring millions of tests taken each year. The machine was developed specifically to score multiple choice items that were used in the 'new-type tests' of the day. The new-type test design had been imported from educational testing into language testing during the First World War, for the construction of the army Alpha tests. In the 1920s new-type tests were taken up in school assessment in order to cope with the rapid expansion of schooling provision across the United States. They could be produced and administered efficiently on the industrial scale that was now required (see Spolsky, 1995: 33–51), and automatic marking made possible by the IBM 805 ensured that the multiple choice item would remain the bedrock of educational testing until the present day.

The practical need to assess large numbers of people cheaply and efficiently, and the advent of the technology to achieve this, sat happily alongside the theoretical concerns of testing and assessment specialists. It had long been known that there was error associated with test scores (Edgeworth, 1888; 1890), and the foundations of educational

testing and assessment were built upon the concept of reliability, developed at the beginning of the century by researchers like Thorndike (1904) and Spearman (1907; 1913). The statistical tools were supplemented in the 1920s with the development of methods to analyse the properties of test items. There was confidence that the new-type tests provided a practical solution to problems in educational assessment, especially the theoretical (and ethical) problems of reliability. Wood (1927) was among the first to undertake the task of reducing the element of chance in taking tests in large-scale language assessment. He was especially concerned with producing tests that would help place students in homogenous classes in New York schools to avoid the learning and teaching problems that existed at the time. Or, as he put it, to avoid having unhappy students in classes where making progress in language learning was a matter of pure chance.

There was a great deal of debate concerning the value of the new-type tests and the limitations of what can be tested using multiple choice items throughout the 1920s and 1930s (Barnwell, 1996: 49). The debate still continues, although those who weigh up the pros and cons of objective item types are all too frequently unaware of the fact that today's arguments merely rehearse those conducted 80 years ago. What is important to recognise now is that multiple choice items (and tests constructed of objective items) are not inherently more reliable than other item or task types. Lado's (1961) important distinction between reliability and 'scorability' is very important: it is possible to have poorly constructed, unreliable, but easily scored multiple choice items, and well constructed, reliable, but time consuming speaking tests. The reason why multiple choice items have been used so much for a century is not some innate reliability of the item type, but the simple fact that it can be scored so easily by a machine.

As we will see later, the scorability of tests by machines is still an area of concern and active research in the current generation of computer-based tests (CBTs). This is true for constructed responses as well as responses to objective items, for the same reasons of cost and efficiency (Burstein, 1997).

From scoring to testing

Since the first book on computerised language testing was published (Holtzman, 1970) the literature on computer-based testing has grown exponentially. Not surprisingly, one major theme has been the rapid change in computer technology that has allowed us to move away from simply scoring tests to developing comprehensive language testing software packages. From a scoring machine in 1935 to the powerful, cheap desktop machines we have today, language testers have been continuously trying to put the available technology to use. Bunderson, Inouye and Olsen, (1989: 367–368) wrote:

> The computer revolution has been marked by the growth in power and sophistication of computing resources. The computing power of yesterday's mainframes is routinely surpassed by today's supermicros. Yesterday's ENIAC computer, which filled an entire room, was less powerful than the current generation of microcomputers, which fit on a desktop

It is only when placing this next to another quotation from the same paper (Bunderson et al. 1989: 371) that we realise how fast technology is in fact changing:

> The memory capacity of most modern delivery systems is evolving rapidly. Most microcomputer workstations now have half to two megabytes of random access memory. Future workstations will use even larger amounts of random access memory. The early, expensive, mass-storage devices are being replaced by inexpensive, high-density, magnetic and opto-electronic devices. Hard-disk storage exceeding 100 megabytes per workstation is becoming more common.

Any discussion of hardware, memory size, and processing capacity will inevitably be out of date within a few years, if not months, of being written. We must simply assume that the constant development, availability and use of larger and faster computers will be the norm for the foreseeable future.

What are the uses to which language testers have put this technology? In a recent review of the role of computers in language testing, Burstein, Frase, Ginther and Grant (1996), isolated eight areas in which computers are now used in language testing. These may be summarised as:

- Test design: the exchange of written and graphic materials between test designers who may be working in different locations.
- Test construction: including item trials, the main function of the computer is envisaged to be the exchange of written and graphic materials, as well as items, between those involved in item writing and revision.
- Item tryout: Items are delivered in their near-final format, responses are stored on the computer, and item level statistics calculated and stored in a database (itembank) with the test items.
- Test item delivery: the delivery of actual tests from databases, including the collection and storage of responses. Appropriate technologies for test-taker identification should also be considered in this phase of the process.
- Item management: storing and updating item information.
- Item scoring and transforming item responses into test scores.
- Item analysis and interpretation: relating the score to some general interpretative scheme for the score.
- Score reporting: delivering scores and related information.

The authors lamented the fact that the use of computers in language testing had not resulted in the creative use of multimedia elements in the way that has happened in instructional programs. And at the time of writing this paper, there is still little in the way of innovative computer-based tests that contain a creative mix of media. This is because, as Frase (1997: 519) points out: 'the obstacles to the successful use of technology for language testing now seem less technical than conceptual.' We turn briefly to some of these conceptual issues, before considering approaches to computer-based testing.

Concepts, constructs and equity

The computer is ideally suited to the delivery of objective items, particularly multiple choice items. As we have shown, when the first computers that could deliver tests became available, existing language testing theory was compatible with the new technology. Classical Test Theory (CTT) and the statistical tools associated with it were developed to

heighten the reliability of objective tests. The first computer-based tests were simply paper and pencil tests that had been designed, constructed and piloted using the tools of CTT. These tests were delivered through the new electronic medium for ease of test administration, delivery and scoring (Alderson, 1988), thus reducing costs even further than had been possible with scoring machines. Recent complex computer adaptive tests (see below for a definition and discussion of CATs) are based on Item Response Theory (IRT) but IRT (and its associated statistical tools) was also developed for the analysis of objective tests.[1] It is therefore not surprising that objective items and contexts that look and feel as if they have been derived from pencil and paper tests continue to appear in computer-based language tests. However, there are other more important reasons why language test designers have been slow to take advantage of the multimedia capability of the computer.

The concerns that language testers have with computer-based tests are not dissimilar from those they have with paper and pencil tests. The most important of these is knowing what the test measures, of the underlying test construct. The ability to make valid inferences from test scores depends upon providing rationales and empirical evidence to support construct validity (Messick, 1989). The introduction of multimedia to a listening test may change the nature of the construct being measured. It is possible that video content changes the process of comprehending listening texts in ways that we do not yet fully understand. Extensive research has not been conducted to discover if the meaning of a score may change because of the visual clues of the medium (Ginther and Chawla, 1997), or whether changes in score meaning are related to test construct or error. Until this research has been done it seems unlikely that test providers such as Educational Testing Services (ETS) will introduce multimedia into the computer-based TOEFL. Construct validity, as an all-inclusive concept, is central to all language testing research. The focus on technology in computer-based and computer adaptive testing is now giving way to the requirements that construct validity be investigated on an on-going basis, as can be seen from the range of papers in a recently published volume on the computer adaptive testing of reading (Chalhoub-Deville, 1999; see also Fulcher, in press for the significance of this volume).

A further major concern is the ethical aspect of computer-based testing. A computer-based test (which is designed to measure the same construct as a paper and pencil based test) should rank order test-takers in approximately the same way as a pencil and paper form of the test, and the two forms should have similar means and standard deviations (APA, 1986). Together these requirements constitute the principle of equivalent forms. Fulcher (1999) reviewed evidence regarding the equivalence of forms, but also considered it important to investigate whether other factors impacted on test scores, such as the test-taker's previous experience with a range of computer uses, their attitudes to technology and taking tests on computers, and whether factors such as age, gender, educational background or L1 would be likely to affect the score on a computer-based test. The issue of the impact of familiarity with computers was also of major concern to ETS prior to the introduction of the computer-based TOEFL in 1998. Kirsch et al. (1998), Eignor et al. (1998), Taylor et al. (1998) and Taylor et al. (1999) investigated the impact of the new delivery medium and found that 16% of test-takers were affected. The solution was to introduce a compulsory tutorial that all test-takers must take immediately before they do

the TOEFL CBT. ETS has taken other steps to familiarise potential test-takers with the format and medium, such as the sampler CD, which is available free of charge. In the writing component of the new TOEFL, there is also the option to type the response or answer in long hand, an option designed to 'bias for the best' by letting test-takers select the medium in which they think they will perform better.

The latest edition of the Standards for Educational and Psychological Testing (1999) raises concerns with regard to computer-based tests, which researchers must address in the future. The following standards are particularly relevant to computer-based tests.

- Standard 2.8: In computer-based tests there is a worry that if the test is speeded, there may be a large impact upon the test score, especially if the test is adaptive and the test-taker responds randomly to items towards the end of the test.
- Standard 5.5: Test-takers should be given an opportunity to respond to sample test items (and their responses monitored) unless they are familiar with the equipment and response type already.
- Standard 6.11: If a test can be taken on computer and in paper and pencil format the interchangeability of the scores should be investigated and reported.
- Standard 4.10 and 8.3: Where a test-taker is offered an alternative test form (such as paper and pencil or computer-based form), there should be enough information available about the characteristics of the two forms to allow the test-taker to make an informed decision.

With specific reference to computer adaptive tests:

- Standard 3.12: Technical manuals for CATs should provide information on the procedures for selecting items, the criteria for selecting the starting point and termination point of a test, for scoring, and for controlling for item exposure.
- Standard 13.18: 'Documentation of design, models, scoring algorithms, and methods for scoring and classifying should be provided for tests administered and scored using multimedia or computers. Construct-irrelevant variance pertinent to computer-based testing and the use of other media in testing, such as the test-taker's familiarity with technology and the test format, should be addressed in their design and use.'

It is clear that the further development of computer adaptive tests is not simply a matter of exploiting the power of the computer as quickly as possible, but of constructing a systematic research agenda that investigates the issues involved. It may be for this reason that, while graphics have been introduced into the TOEFL CBT (ETS, 1998), the TOEFL test specifications stop short of using the full multimedia potential of the computer.

Computer adaptive testing

Although any test that is currently delivered using paper and pencil can also be delivered by computer, the most important development of the last decade has been computer adaptive testing, in which the computer branches to certain sub-tests (branching routines) or selects the next test item (adaptive routines) depending upon the response pattern of the individual test-taker. The first CATs were developed in the 1970s (Gruba and Corbel, 1998; Dunkel, 1999: 80) although it is not until recently that they have come into

widespread use, generating much research and comment.[2] CATs have been made possible by the extensive use of Item Response Theory, and the development of algorithms that drive the test program to select and deliver test items, score responses, and provide immediate feedback to test-takers. It is beyond the scope of this paper to provide an introduction to IRT methods, but there are a number of excellent texts available for the reader who wishes to learn more about the measurement theory underlying computer adaptive testing.[3]

Bunderson et al. (1989: 381) describe the development of CATs as the second generation of computerised testing, capable of adapting test content on the fly to suit the estimated ability of the test-taker. One of the most advanced CAT systems is *FastTEST*, produced by Assessment Systems Corporation in 1999 as a successor to *MicroCAT*, which was first released in 1984.[4]

Computer adaptive tests constructed and delivered with systems like *FastTEST* have a number of major advantages. Firstly, all items and item level information is contained in an item bank on the local machine or network. Attached to each item is information that is used by the program's algorithm in selecting items or subsets of items for delivery. The information contains item statistics like difficulty, discrimination, and perhaps a guessing parameter. It may also contain information on content, context, or any other tag that would be relevant to item selection for specific purposes testing. This means that test items can be selected to individualise the test by matching it to the test-takers' needs or the requirements of score users. Secondly, the selection of the next item or sub-test is dependent upon the responses of the test-taker. The algorithm may select a more difficult item for learners who get the responses to previous questions correct, and easier items for learners who answer many items incorrectly. Therefore, no test-taker takes exactly the same test as any other, assuming that the item bank is reasonably large. This increases test security. Thirdly, the number of items that the test-taker is required to attempt is reduced, as the computer will terminate the test once an assessment of the test-taker's ability level has been estimated within pre-set error (or other relevant) parameters. Not only does this save time and resources in terms of test delivery and the amount of time needed to administer tests, it also provides instant results and reporting. Taking an adaptive test can therefore be more motivating for many learners. Finally, if a large enough bank of items with very high or low facility values can be constructed, it should in principle be possible to identify students who are extremely able, and those who have very low ability. This is not possible with non-adaptive tests because the test would need to be exceptionally long and contain items with facility values that cover the entire ability range.

However, there are a number of disadvantages associated with CATs. CATs can only work if there is a large number of items in an item bank, which are calibrated to a measurement scale constructed using Item Response Theory. Building up a sufficiently large item bank can be time-consuming and costly. The more parameters the algorithm uses in CAT administration, the larger the sample size needed for pre-testing and calibration. For example, when using a three-parameter model item statistics are unstable using samples of less than 1000 test-takers. If the item bank is not sufficiently large, with sufficient items across the entire ability range of the test-takers, there may be item overexposure (threatening security) or a failure to adequately estimate the ability of very

able or very weak students. While the use of calibrated items from a bank allows clearer interpretation of score meaning, achieving this should not be seen as an easy process. Establishing criterion-referenced meaning at various points on a scale, especially for cut scores, requires careful research. This is especially true in CATs where the test-takers are not taking the same test. The third problem associated with the item bank is one of sampling. It is frequently assumed that items are written to adequately reflect the domain to which the score user wishes to make inferences, and the implementation of a CAT means that only a small proportion of the items are selected for any individual test. It is therefore appropriate that the issue of content validity of a CAT is problematised, so that test developers consider whether, and to what degree, the CAT should be forced to include a representative sample of items in the test, even if they are not needed in order to place a student on the ability scale. Finally, in a CAT the test-taker is not allowed to omit items. The reason for this is simply that if learners respond only to items that they think they are going to get right, the ability estimate will be unnaturally high. All items must be answered for the computer to estimate ability reliably, unlike paper and pencil tests where the test-takers have the opportunity to miss items if they wish, and review items if they have time at the end of the test. In CATs, this freedom is removed. As yet, however, there has been no research to suggest that this is demotivating or disadvantageous for any identifiable groups of test-takers.

It can be seen that CATs have major advantages over paper and pencil tests, despite the research that still needs to be conducted before the full potential of the computer can be harnessed in test design. Nevertheless, there is one other issue that requires consideration. In some countries (mainly in the United States) there is legislation requiring the disclosure of test papers (see Brown, 1997: 53). In the United Kingdom, the Department of Education has recently considered introducing legislation to force examination boards to return all examination scripts to students after they have been scored. CATs rely on large secure item banks that are expensive to build and maintain. If the item banks must be periodically disclosed to test-takers, then CATs would become prohibitively expensive. Test developers would be forced back to first generation non-adaptive computer-based tests. While legislators in the United States are aware of these problems and considering disclosure laws in the light of modern test theory and CAT developments, this is not the case in other countries such as the United Kingdom, where there is little awareness of measurement issues.

Despite the growing commercial availability of CAT software with user friendly interfaces that require only a passing knowledge of Windows operating systems to use, teachers and language teaching institutions should beware of moving from conventional testing to computer adaptive testing. Without the resources and expertise to develop and operate the systems, it is better to remain with good conventional tests that produce better quality information within local settings.

Testing on the Internet

Although the Internet, and the World Wide Web in particular, is a global information distribution network that would easily allow the delivery of tests anywhere in the world, its potential has not yet been realised. Nevertheless, the range and variety of Internet-

based tests is growing. Links to tests currently available are maintained at the Resources in Language Testing Page, which is frequently updated.[5]

The interactivity that is currently available on the Web is provided by programs stored on the server written in *PERL* script (Practical Extraction and Report Language) or downloaded to the client's machine in *Java*. This allows computer-based tests or CATs to be scored on-line. However, at the time of writing this paper, all on-line tests are available only as low-stakes 'quizzes.' This is mainly because large scale high-stakes test delivery over the Web faces serious security problems. Until the security issues associated with the transfer of information over the Internet have been solved, it is unlikely that testing organisations will use the Web, and will prefer to use third-party computer installations such as those provided by Sylvan.[6] Whilst testing on the web remains non-commercial, it is unlikely that any significant CATs with large item banks will be available within the near future. One exception to this may be DIALANG.[7] Funded by the European Union, the DIALANG project was designed to produce diagnostic tests in 14 European languages delivered as a CAT over the Web (see Alderson, in press). Although not adaptive at item level, the system allows branching routines depending upon an initial self-assessment, and on-going estimation of test-taker ability. When the program is released it is expected to provide feedback to the test-taker on the relationship between self-assessment and estimated ability, and benchmark the estimated ability to Council of Europe proficiency level descriptors.

Until large testing organisations such as ETS are able to utilise the Internet for high-stakes test delivery, first generation computer-based tests will remain a very real option for language testers. The delivery of these traditional CBTs on the Internet is of particular interest, for a variety of reasons.

Firstly, the only software needed to take the tests is a standard browser. These can be loaded onto any type of computer, making the test delivery system truly platform independent. The only requirements relate to hardware (the need for a modem), and a reasonably fast processor to download the information from the host server providing the test. The second important advantage of the Internet as a means of delivery is that the tests can be delivered to any machine linked to the Internet, at any time convenient to the provider and the client. In distance learning programmes such arrangements can be beneficial to both the learner and the tutors.

The Web also provides advantages in the flexibility of test design without the need to resort to third-party plug-ins. It is quite feasible, for example, to use the frames facility of the browser to divide the computer screen into windows, each of which contains a content page. Prompts may be set up on a series of frames that incorporate text, images, audio, and video, where computer links are reliable and quick. In fact, the flexibility of html in designing web pages makes it possible to design a range of novel task types through the imaginative combination of multimedia in a frames environment for low-stakes testing or research (Fulcher, 1998). A further advantage of delivering tests over the Web is that links can be established to information, help facilities, databases, or libraries, to deliver the kind of indirect performance test frequently recommended for placement purposes in academic programmes (see Robinson and Ross, 1996). Tests need no longer be self-contained, watertight units, but involve the use of information from the outside world, to any degree the test designer wishes to incorporate it. This potential can be used to increase the 'authenticity' of some testing activities.

In computer-based testing on the Internet, innovation is possible where there is flexibility over the format and content of the prompt. However, it is not as easy to be as innovative in the area of item type, as we have indicated above. Most Internet browsers support multiple choice, multi-choice, pull-down menu and constructed response item types, and combinations of these. For example, multiple pull-down menus can provide matching or sequencing items. Constructed response items may be of two types: limited constructed response where a word or short phrase is required, and which is automatically scored against a template, and extended constructed response, which must be e-mailed to human raters for scoring. In this respect, little has changed since Alderson (1988) found it difficult to design innovative item types for computer-based language tests.

From a measurement perspective, Internet testing raises many questions that still need to be investigated, as we have argued above. At present, there is not enough research evidence to justify introducing the novelty of what can now be achieved, except for use in low-stakes testing and research. Measurement and ethical questions must be addressed in relation to the Internet, just as they must in developing CAT listening tests with video instead of audio. In summary, we are currently in a situation where innovation and flexibility are possible but implementation is not a pressing concern until the conceptual problems associated with the new medium have been thoroughly researched.

Testing, artificial intelligence and constructed response

The third and fourth generation of computerised testing, as described by Bunderson et al. (1989), whilst visionary, are nevertheless still some way in the future. The third generation of computerised testing is the continuous assessment of learning and the projection of learning trajectories from the current ability level of the student to another ability level at some point in the future. The assumption is that it is possible to calculate trajectories in language learning in a meaningful way. Given what we currently know about language acquisition, including U-shaped and discontinuous learning (Larsen-Freeman and Long, 1991: 105–107; Perkins et al., 1996), and taking into account the multitude of variables that affect language learning, it seems unlikely that the progress of individuals can be meaningfully predicted very far into the future.

This makes it more unlikely that we will see the development of the predicted fourth generation of assessment (Bunderson et al. 1989: 398–402), in which artificial intelligence will be brought to bear on continuous measurement in order to provide advice on learning style and content selection for the learner, related to the current estimated stage of learning. This fourth generation of tests would have all the properties of the third generation, but would be linked to expert second language acquisition systems. The field of second language acquisition is currently not able to provide such an expert system, and even if a model of language development could be generally agreed, calibrating the test to the theoretical model would be a major project that would occupy researchers for many years.

In recent years there have been a number of significant advances. We review two here, both relating to the scorability of constructed responses.

Bernstein (1997) reports on the development of *PhonePass*™, which is a test of speaking conducted over the telephone with a computer.[8] The testing system relies on

the computer being able to match the pronunciation of the speaker on any given word using a statistical model derived from a large database of (intelligible) native speakers of American English, and evaluating the rate of delivery. The test takes ten minutes, and diagnostic sub-scores are returned for reciting/pronunciation, reading fluency, repeat accuracy, repeat fluency, and listening vocabulary. Ordinate, the company that has developed *PhonePass*, is unusual among private testing companies (and, for that matter, public examination boards in the UK!) in that it has conducted extensive research into the reliability and validity of the system it has developed. Ordinate reports reliability coefficients for *PhonePass* that are comparable (and sometimes higher) than human raters, and overall correlations between scores awarded by human judges and *PhonePass* of .93. Ordinate has also commissioned a number of studies to investigate its reliability and validity, and made these available on the Internet in portable document format (Bernstein, 1998; 1999a; 1999b).

The problem with a test like *PhonePass* is that it relies for its validity on the correlation with direct measures such as the Oral Proficiency Interview (OPI). It is possible that an estimate of speech rate would correlate with a direct test of speaking, just as it is possible that the height of a learner in Spain would correlate with a measure of vocabulary size. These factors are related through other variables. Yet, it would be difficult to claim that measuring height and making an inference about language ability had any construct validity. The definition of 'fluency' in *PhonePass* (Bernstein, 1998) as rhythm, phrasing and pausing is far from the complex applied linguistic notions that are used in non-computerised test development (Fulcher, 1996). Nevertheless, it is work like that of Bernstein that will lead to continued improvement in speech recognition technology. This may ultimately lead to a new generation of automated speaking tests that could support stronger construct claims. In the meantime, Ordinate can offer a cheap on-demand test that is certainly predictive of speaking ability, even if its construct validity may be questioned.

Another exciting development concerns the e-rater,[9] developed by ETS primarily for use on Graduate Management Admission Test (GMAT) and the Graduate Record Examinations (GRE) to rate essays. Research has also been conducted into the ability of e-rater to automatically score writing from non-native speakers of English using scripts from the Test of Written English (TWE). Burstein and Chodorow (1999) report agreement between e-rater and human graders in 92% of cases. E-rater automatically builds models for individual writing prompts using large numbers of human scored writing samples, based on 52 syntactic, discourse and topic variables. It then matches new writing samples against the model to produce the score. Initial research shows that e-rater performs differentially across language groups, and future research at ETS is likely to concentrate on whether different models are needed for different language groups. It should of course be stressed that e-rater is a research project. It is being used operationally only in conjunction with human raters, and is not being used operationally at all with non-native writers.

In these two examples, we can see the issue of scorability – this time with constructed response tests. While *PhonePass* can be criticised for relying heavily on correlational evidence for its validity, the e-rater represents a contribution to construct research. The advances in natural language processing since the reviews of Freedle (1990) mean that

the process of machine rating has started to become much more like the process of human rating in the consideration of linguistic elements of the text. It is fair to say that the e-rater represents the first application of artificial intelligence that may successfully help with construct definition.

Communication and the dissemination of research

It may be noticed that many of the sources cited in this chapter are available on-line. Perhaps one of the enduring benefits of computers to language testing will be the speed and ease of communication, and the availability of information. This is achieved through discussion lists dedicated to language testing issues and web sites devoted to the transmission of language testing and measurement information.

LTEST-L remains the active open discussion list of testing issues. ILTA-MEM is the list of the International Language Testing Association (ILTA), and is only open to members. Together, these lists allow the speedy transmission of information, and keep teachers and researchers in touch with the latest developments in the field.

On the Web, sites such as *Assessment and Evaluation on the Internet* (http://ericae.net/nintbod.htm) and the *Resources in Language Testing Page* (http://www.surrey.ac.uk/ELI/ltr.html) provide electronic focal points for the dissemination of information. The latter, now receiving reviews in text publications such as Sperling (1998) and Douglas (2000), provides an electronic hub for language testing in particular. The Home Page of ILTA is maintained at the same site, which has recently been expanded to contain a series of introductory videos on language testing for language teachers (Fulcher and Thrasher, on-line). This novel approach to dissemination of language testing information is part of a strategy to achieve the educational objectives of ILTA. The page is now being used in introductory language testing courses for teachers in South America, Europe and the Far East.

Test-takers also have access to information about the tests they may have to take, and can download sample papers and advice on test-taking. The University of Cambridge Local Examinations Syndicate maintains a download page with handbooks and sample papers for most of its tests (http://www.cambridge-efl.org/support/dloads/index.html). ETS offers the same for TOEFL, but adds information on test-taking for learners with disabilities, tutorials, a download library, order forms for free sample disks, and a store to purchase preparation materials (http://www.toefl.org/).

Researchers, teachers and test-takers therefore have more information on language testing available than ever before, and this would not have been possible without Internet communication.

Conclusions

The use of computers in language testing, and the use of computer adaptive tests, has gone beyond the stage of discussing how technology can lead to innovation in design that was prevalent even a few years ago. Nevertheless, as long as scorability remains an issue in large testing programs where cost is an important factor, we will continue to see researchers trial any solution that automates the process.

The real focus in the coming years will be on conducting research that addresses the

fundamental questions that must be asked about any test, any testing procedure or system, whether paper- or computer-based: what inferences can be drawn from the test scores. Some scorability research, like that into the e-rater, will help here, because the researchers are trying to copy in the program what human raters respond to in text and, in the process, help clarify the construct of writing ability. The issue that will dominate discussion and research on computers in language testing in the next decade, as Chalhoub-Deville (1999: x) rightly predicts, will be test construct.

Notes

[1] For a clear explanation of CTT and IRT and their associated statistical tools see Davidson F (in press) 'The Language Tester's Statistical Toolbox'. In Fulcher G (Ed.) Expanding perspectives on language testing in the 21st century. Special Edition of *System 28 4*

[2] For recent technical reviews of the potential advantages and disadvantages of computer adaptive testing see Brown, 1997; Chalhoub-Deville et al, 1997; Chalhoub-Deville & Deville, 1999; Dunkel, 1997; 1999; Fulcher, 1998

[3] For a brief and relatively non-technical introduction to IRT see Baker, 1997. A much more detailed text is Crocker & Algina, 1986. For readers who wish to understand the statistics involved in IRT a useful introduction is provided by Henning, 1987

[4] A demo version of *FastTEST* can be downloaded from the ASC website at http://www.assess.com/FastTESTPro.html

[5] The Resources in Language Testing Page is available at http://www.surrey.ac.uk/ELI/ltr.html

[6] Sylvan Psychometric is contracted by ETS to deliver the TOEFL CBT worldwide. Their services and operations are described on their web page at http://www.sylvanprometric.com/

[7] DIALANG is an acronym for Diagnostic Language Assessment Details of the DIALANG project are available from the project website at http://www.jyu.fi/DIALANG/

[8] *PhonePass*™ is the trade mark of the computerised speaking test produced by Ordinate INC. Details of Ordinate products and services is available from its web site at http://www.ordinate.com/index.html. Visitors may arrange to take a sample test over the Internet

[9] Full-text research papers on e-rater are made available by ETS at http://www.ets.org/research/erater.html

Bibliography

AERA APA NCME (1999) *Standards for Educational and Psychological Testing*: Washington DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education

Alderson J C (1988) Innovations in Language Testing: Can the Microcomputer Help?: *Special Report No 1 Language Testing Update* University of Lancaster

Alderson J C (In press) Technology in Testing: The present and the future in Fulcher G (Ed.) *Expanding perspectives on language testing in the 21st century* Special Edition of System on Language Testing *System 28*

APA (1986) *Guidelines for Computer Based Tests and Interpretations*: American Psychological Association

Baker R (1997) Classical Test Theory and Item Response Theory in Test Analysis: *Special Report No 2: Language Testing Update* University of Lancaster

Barnwell D (1996) *A History of Foreign Language Testing in the United States* Arizona Bilingual Press

Bernstein J (1997) Speech Recognition in Language Testing. In Huhta A, V Kohonen, L Lurki-Suonio & S Luoma (Eds.) *Current Developments and Alternatives in Language Assessment* Jyvaskyla University

Bernstein J (1998) *Construct Comparison between the Language Proficiency Interview (LPI) and the PhonePass™ Test* Available on-line at http://www.ordinate.com/pdf/ConstructComparisonLPI990826.pdf

Bernstein J (1999a) *PhonePass™ Testing: Structure and Construct* Available on-line at http://www.ordinate.com/pdf/StructureAndConstruct990826.pdf

Bernstein J (1999b) *PhonePass™ Data Analysis: Correspondence with Oral Interviews and First-Language Bias Analysis* Available on-line at: http://www.ordinate.com/pdf/StructureAndConstruct990826.pdf

Brown J D (1997) Computers in Language Testing: Present research and some future directions *Language Learning & Technology 1 1* Available on-line at http://polyglot.cal.msu.edu/llt/vol1num1/brown/default.html

Bunderson C V, D I Inouye & J B Olsen (1989) The four generations of computerized educational measurement. In Linn R L (Ed.) *Educational Measurement (3rd edition)* American Council on Education

Burstein Jill & Martin Chodorow (1999) Automated Essay Scoring for Nonnative English Speakers *Joint Symposium of the Association of Computational Linguistics and the International Association of Language Learning Technologies Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing College Park* June 1999 Available on-line at http://www.ets.org/research/acl99rev.pdf

Burstein J, L T Frase, A Ginther & L Grant (1996) Technologies for Language Assessment *Annual Review of Applied Linguistics 16*

Burstein J (1997) Scoring Rubrics: Using Linguistic Description to Automatically Score Free-Responses. In Huhta A, V Kohonen, L Lurki-Suonio & S Luoma (Eds.) *Current Developments and Alternatives in Language Assessment* Finland Jyvaskyla University

Chalhoub-Deville M (1999) (Ed.) *Issues in computer adaptive testing of reading proficiency Issues in Language Testing Vol 10* Cambridge University Press

Chalhoub-Deville M Alcaya C & Lozier V M (1997) *Language and Measurement Issues in Developing Computer-Adaptive Tests of Reading Ability: The Minnesota Approach.* In Huhta A, V Kohonen, L Lurki-Suonio & S Luoma (Eds.) *Current Developments and Alternatives in Language Assessment* Jyvaskyla University

Chalhoub-Deville & Deville C (1999) Computer adaptive testing in second language contexts *Annual Review of Applied Linguistics 19*

Crocker L & J Algina (1986) *Introduction to Classical and Modern Test Theory* Holt Rinehart & Winston

Davidson F (in press) The Language Tester's Statistical Toolbox in Fulcher G (Ed.) *Expanding perspectives on language testing in the 21st century* Special Edition of System on Language Testing *System 28*

Douglas D (2000) *Assessing Languages for Specific Purposes* Cambridge University Press

Dunkel P A (1997) Computer-Adaptive Testing of Listening Comprehension: A Blueprint for CAT Development. *The Language Teacher Online 21 10* Available on-line at http://langue.hyper.chubu.ac.jp/jalt/pub/tlt/97/oct/dunkel.html

Dunkel P A (1999) Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning & Technology 2 2* 77–93. Available on-line at http://polyglot.cal.msu.edu/llt/vol2num2/article4/index.html

Edgeworth F Y (1888) The statistics of examinations *Journal of the Royal Statistical Society 51*

Edgeworth F Y (1890) The element of chance in competitive examinations *Journal of the Royal Statistical Society 53*

Eignor D, C Taylor, I Kirsche & J Jamieson (1998) Development of a Scale for Assessing the Level of Computer Familiarity of TOEFL Examinees. *TOEFL Research Report 60* Princeton NJ: Educational Testing Service. Available on-line at ftp://etsis1.ets.org/pub/toefl/275756.pdf

ETS (1998) *Computer-based TOEFL score user guide* Educational Testing Service

Frase L T (1997) Technology for Language Assessment and Learning: Introduction and comments on the State of the Art. In Huhta A, V Kohonen, L Lurki-Suonio & S Luoma (Eds.) *Current Developments and Alternatives in Language Assessment* Jyvaskyla University

Freedle R (1990) *Artificial Intelligence and the Future of Testing* Lawrence Erlbaum

Fulcher G (1996) Does thick description lead to smart tests? A data-based approach to rating scale development *Language Testing 13 2*

Fulcher G (1998) Computer based language testing: The call of the internet. In Coombe A (Ed.) *Current Trends in English Language Testing Vol 1 Conference Proceedings for CTELT 1997 and 1998* Al Ain United Arab Emirates TESOL Arabia

Fulcher G (1999) Computerizing an English Language Placement Test *English Language Teaching Journal 53 4*

Fulcher G (In press) Review of Chalhoub-Deville M 1999 (Ed.) *Issues in computer adaptive testing of reading proficiency*: Issues in Language Testing Vol 10 Cambridge University Press *Language Testing 17 2*

Fulcher G & Thrasher R *Video FAQs: Introducing Topics in Language Testing* ILTA [online] available: http://www.surrey.ac.uk/ELI/ilta/faqs/main.html

Ginther A & Chawla A (1997) Multimedia – Words with Pictures: Unpacking the Effects of Visual Accompaniments to Listening Comprehension Items. In Huhta A, V Kohonen, L Lurki-Suonio & S Luoma (Eds.) *Current Developments and Alternatives in Language Assessment* Jyvaskyla University

Gruba P & C Corbel (1998) Computer-based testing in Clapham C & Corson D (Eds.) Language Testing and Assessment Vol 7 *Encyclopedia of Language and Education* Dordrecht: Kluwer Academic Publishers

Henning G (1987) *A Guide to Language Testing: Development evaluation research* Newbury House

Holtzman W H (Ed.) (1970) *Computer-assisted instruction testing and guidance* Harper Row

Kirsch I, J Jamieson, C Taylor & D Eignor (1998) Computer Familiarity Among TOEFL Examinees *TOEFL Research Report 59* Educational Testing Service. Available on-line at ftp://etsis1.ets.org/pub/toefl/275755.pdf

Lado R (1961) *Language Testing* Longman

Larsen-Freeman D & M Long (1991) *An Introduction to Second Language Acquisition Research* Longman

Messick S (1989) Validity. In Linn R L *Educational Measurement* American Council on Education/Macmillan 1

Perkins K, S R Brutten & S M Gass (1996) An investigation of patterns of discontinuous learning: Implications for ESL measurement *Language Testing 13 1*

Robinson P & S Ross (1996) The Development of Task-Based Assessment in English for Academic Purposes Programs *Applied Linguistics 17 4*

Spearman C (1907) Demonstration of formulae for true measurement of correlation *American Journal of Psychology 18*

Spearman C (1913) Correlations of sums and differences *British Journal of Psychology 5*

Sperling D (1998) *The Internet Guide for English Language Teachers* Prentice Hall

Spolsky B (1995) *Measured Words* Oxford University Press

Taylor C, J Jamieson, D Eignor & I Kirsch (1998) *The Relationship between Computer Familiarity and Performance on Computer-based TOEFL Test Tasks TOEFL Research Reports 61* Princeton NJ Educational Testing Service. Available on-line ftp://etsis1.ets.org/pub/toefl/275757.pdf

Taylor C, I Kirsch, D Eignor & J Jamieson (1999) Examining the Relationship Between Computer Familiarity and Performance on Computer-based Language Tasks *Language Learning 49 2*

Thorndike E L (1904) *An Introduction to the Theory of Mental and Social Measurements* Science Press

Wood B D (1927) *New York experiments with new-type modern language tests* Macmillan