Research Journal of English Language and Literature (RJELAL)
'eer Reviewed (Refereed) International Journalhttp://www.rjelal.com;
Email:editorrjelal@gmail.com ISSN:2395-2636 (P); 2321-3108(O)

Vol.5.Issue 2. 2017
(April-June)

REVIEW ARTICLE

INTERNATIONAL STANDARD SERIAL NUMBER INDIA
2395-2636 (Print):2321-3108 (online)

Computer Assisted Language Testing
Benefits, Constraints and Future Directions

MATIN RAMAK[1], SINA RAHMATTALAB ZIABARI[2]
[1]M.A,Instructor of ELT, Azad University, Iran
[2]M.A,Instructor in Law, Azad University, Iran
Email: m.ramak90@yahoo.com[1];rahmattalab@yahoo.com[2]

MATIN RAMAK

SINA RAHMATTALAB ZIABARI

ABSTRACT

Assessment of learners' language ability is an important part of language education, which has been affected by computer technology at least as significantly as language learning has. Computer-assisted language assessment (CALA) employs the use of technology to facilitate, contextualize, and enhance the assessment of linguistic abilities. CALA is becoming normalized, concomitant with advances in technology and its propagation in language learning contexts. Within CALA, though, most attention has been devoted to technology in language tests (computer-assisted language testing or CALT).In this paper after a brief overview of computer assisted testing (CAT) in general, recent developments in the use of computers in language testing in two areas: (a) item banking (b) computerized-adaptive language testing is discussed.This paper also attempts to showbenefits, problems/constraints of applying computers in language testing .Therefore, authorities should be aware of the positive and negative aspects of a computer adaptive test when they want to administer a test in adaptive format. Moreover this article provides information about future direction of CALT to those interested in utilizing computers in language testing.

Keywords: Computer Assisted Language Assessment, Item Banking, Computer Adaptive Testing, Computer Assisted Language Testing

1. Introduction

Assessment of learners' language ability is an important part of language education, which has been affected by computer technology at least as significantly as language learning has. Computer-assisted testing is an assessment model in which candidates or test takers answer questions or complete exercises that are part of a computer program. In many cases, computer tests also include automatic scoring. This occurs when there are a finite number of correct answers, such as in multiple choice testing models. When short answer and essay questions are included in computer-assisted testing, a grader normally reads answers and enters grades into a database. Computer-assisted testing is used for standardized tests, for psychological and skill assessment, in classrooms, and may even be used by individuals who wish to test themselves.

Proponents of computer-assisted testing believe that it makes recording scores much easier for scorers and instructors. Individuals who take these exams often can receive their scores immediately. Some critics, however, believe that people with different ways of learning and processing information may find computer testing difficult. Computer assisted assessment is often

KY Publications
www.kypublications.com

largely associated with knowledge related skills instead of the overall conceptualization of topics. It can also be difficult to have an objective set of questions that effectively represent the information within the teaching curriculum. One of the major benefits provided by computer assisted assessments is the ability to individually target student skill sets with a minimal time investment. This allows for more time to be spent teaching and learning and far less time spent in the assessment process for both teachers and students. The initial implementation of computer assisted assessments can be met with some serious upfront costs. However, continued development of computer assisted assessment techniques are beginning to lower some costs, such as programs that are web based rather than static software that requires customization. The benefits of computer assisted assessments continue to expand within education and will likely only continue to grow as technology improves to become more accessible for students and teachers.

2. ComputerAssisted Language Testing

The use of computers and electronic devices has become widespread all around the world; specifically, computers and on-line processes were increasingly used for evaluating the language proficiency of English learners (Fleming &Hiple, 2004). These improvements in computer technologies have affected many parts of educational settings such as learning, testing and assessment (Bennett, 2002; Pommerich, 2004). The use of computer technology in the field of language assessment is referred to as Computer-Assisted Language Assessment or Computer-Assisted Language Testing (CALT), both the terms are used interchangeably.

According to José Noijons (1994), CALT is "an integrated procedure in which language performance is elicited and assessed with the help of a computer (P.38)." Chapelle (2010) distinguishes three main motives for using technology in language testing: efficiency, equivalence, and innovation. Efficiency is achieved through computer-adaptive testing and analysis-based assessment that utilizes automated writing evaluation (AWE) or automated speech evaluation (ASE) systems. Equivalence refers to research on making computerized tests equivalent to paper and pencil tests that are considered to be "the gold standard" in language testing. Innovation—where technology can create a true transformation of language testing—is revealed in the reconceptualization of the L2 ability construct in CALT as "the ability to select and deploy appropriate language through the technologies that are appropriate for a situation" (Chapelle& Douglas, 2006, p. 107).

In addition, innovation is exemplified in the adaptive approach to test design and automatic intelligent feedback provided with the help of AWE and ASE technologies integrated in computerized tests. Early on, computer-based tests of foreign language learning involved item types that were easily scored by a computer. Item types included multiple-choice, multiple-select, drag-and-drop, and short-answer response and were presented linearly as they were on their paper-and pencil counterparts. This led to many comparison studies between computerized and paper-and-pencil versions of the same test, and this research still continues today. Eventually, this format changed. Instead of relying solely on discrete item types, test takers were asked to respond to tasks that were more like real-world tasks. In addition, they were asked to produce more open-ended responses. The challenge has been in scoring such items, both in terms of developing the criteria for scoring and in developing programs to help with scoring, and research along this line continues to flourish, especially in the area of computerized tests of writing ability (Goodfellow, Lamy, and Jones, 2002;Li, 2000). In addition, the field quickly incorporated the adaptive functions of computerized testing.

3. Current State of Knowledge on Computers in Language Testing

In reviewing the literature on computers in language testing, there aretwo recurring sets of issues; item banking andcomputer adaptive testing which will be discussed as following:

3.1 Item Banking

Item banking covers any procedures that are used to create, pilot, analyze, store, manage, and select test items so that multiple test forms can be created from subsets of the total "bank" of items. With a large item bank available, new forms of tests can be created whenever they are needed. While the underlying aims of item banking can be accomplished by using traditional item analysis procedures (usually item facility and item discrimination indexes; for a detailed description of these traditional item analysis procedures, see Brown, 1996), a problem often occurs because of differences in abilities among the groups of people who are used in piloting the items, especially when they are compared to the population of students with whom the test is ultimately to be used. However, a relatively new branch of test analysis theory, called item response theory (IRT), eliminates the need to have exactly equivalent groups of students when piloting items because IRT analysis yields estimates of item difficulty and item discrimination that are "sample-free." IRT can also provide "item-free" estimates of students' abilities. Naturally, a full discussion of IRT is beyond the scope of this article.

However, Henning (1987) discusses the topic in terms of the steps involved in item banking for language tests and provides recipe-style descriptions of how to calculate the appropriate IRT statistics.Green (1988) outlines some of the problems that might be encountered in using IRT in general, and Henning (1991) discusses specific problems that may be encountered with the validity of item banking techniques in language testing settings. Another serious limitation of IRT is the large number of students that must be tested before it can responsibly be applied. Typically, IRT is only applicable for full item analysis (that is, for analysis of two or three parameters) when the numbers of students being tested are very large by the standards of most language programs, that is to say, in excess of one thousand. Smaller samples in the hundreds can be used only if the item difficulty parameter is studied.Minimal item banking can be done without computers by using file cards, and, of course, the traditional item analysis statistics can be done (using the sizes of groups typically found in language programs) with no more sophisticated equipment than a hand-held calculator. Naturally, a personal computer can make both item banking and item analysis procedures much easier and much faster. For example, standard database software can be used to do the item banking, (e.g., Microsoft Access, 1996; or Corel Paradox, 1996).

An example of a software program specifically designed for item banking is the PARTest (1990) program. If PARTest is used in conjunction with PARScore (1990) and PARGrade (1990), a completely integrated item banking, test analysis, and record-keeping system can be set up and integrated with a machine scoring system.

3.2 ComputerAdaptive Testing (CAT)

Language performance can be assessed through different procedures. Computerized testing is one of them. Computerized testing originated in the early 1970s (Drasgow, 2002; Wainer, 1990).The emergence of new technologies resulted in development and implementation of computerized testing in large-scale testing programs such as licensure, certification, admissions, and psychological tests (Kim & Huynh, 2007). One of the computerized testing is Computer Adaptive Test (CAT). Specifically, the first (CAT) was created by Larson and Madsen (1985) at Brigham Young University, in the USA. Besides, After Larson and Madsen (1985), several scholars (e.g., Kaya-Carton, Carton &Dandonoli, 1991; Burston&Monville-Burston, 1995; Brown & Iwashita, 1996; Young, Shermis, Brutten& Perkins, 1996) were motivated to construct and develop more computer adapted tests throughout the 1990s. Finally, some standardized tests were administered in computer-adaptive testing format. In 1998, the Test of English as a Foreign Language (TOEFL) started to use computer-adaptive testing format (Mojarrad et al., 2013).

Computer adaptive testing is also called “tailored testing” (Madsen, 1991). Noijons (1994, p. 38) defines adaptive testing as "an integrated procedure in which language performance is elicited and assessed with the help of a computer, consisting of three integrated procedures including: generating the test, interaction with candidate, evaluation of response". Furthermore, Computer-adaptive tests (CATs) are technologically advanced assessment measures (Dunkel, 1999) that have been used in L2 testing since the 1980s. They use sophisticated algorithms to move examinees from one item to the next based on the examinee’s performance on the last item. (Sets or blocks of items used for adaptive purposes are called test lets, and CATs that use them are called semi-adaptive tests). Brown outlined CAT advantages as such: “(a) the items are selected and fitted to the individual students involved, (b) the test is ended when the student’s ability level is located, and, as a consequence, (c) computer adaptive tests are usually relatively short in terms of the number of items involved and the time needed ” (1997, p. 46). This advantage helps with large-scale administrations and keeps test takers from being overburdened by items that are too easy or difficult. Besides, the use of CATs decreases the amount of time needed for test preparation and marking and it can increase consistency of the results (Callear and King 1997).

Additionally, test management is flexible, scores are immediately available, and it may motivate examinees (Linacre, 2000; Rudner, 1998) because items are appropriate for their own level and their test anxiety may be reduced (Mulkern, 1998). Besides, efficiency is the main advantage of CATs (Weiss, 1990; Straetmans&Eggen, 1998). CATs are cost saving compared to conventional methods. In conventional methods, different tests should be given to different groups of students and it is very time consuming to prepare different tests.

Additionally, the use of CATs can be very beneficial, where a large number of learners should be placed into different classes immediately (Weiss, 1990); because through the flexi-level strategy, examinees do not need to answer a large number of question which are too difficult or too easy for them. In fact, in computer adaptive tests, the examinees can be given different tests which are appropriate for their own specific level (Larson and Madsen, 1985). Because each test is adapted to each examinee level, more information can be gathered from computer adaptive tests compared to traditional tests (Young et al., 1996). The last advantages of CATs, is "greater precision of measurement", which causes more accurate mastery classification"(Weiss, 1990, p. 454). This greater precision of measurement is due to using items that are at the maximum discrimination level. Besides, the score of each examinee was determined based on both the percentage of questions which were answered correctly and the difficulty level of these questions. As a result, if two examinees answer the equal percentage of questions correctly, the one who answers more difficult questions gets higher score (Economides and Roupas, 2007).

4. Benefits of CALT

The sophisticated and adaptive nature of computer assisted language testing has many benefits that can be used for overcoming many of the prevailing problems/constraints in the field of traditional testing. Many scholars like Carol A. Chapelle and Dan Douglas (2006), Dandonoli (1989), Larson (1989), Stansfield (1990), Madsen (1986,1991) have advocated the use of computer technology in the field of language assessment and testing due to its benefits. These benefits are seen both from the angle of testing methodology and human considerations.

The first benefit is that it helps to overcome many of the administrative and logistic burdens associated with tradition testing practices by making the test available wherever and whenever the test taker can logon to the internet or can insert a disk into a CD-ROM drive. It also reduces the logistical burdens by transmitting test materials electronically. The use of the internet for test delivery in the form of web based testing or WBT has been the most significant contribution to the field of language assessment to overcome many of these logistical and administrative problems as rightly observed by Roever (2001), an enthusiast of web-based testing, in the following word: “Probably the single biggest logistical advantage of a WBT [web-based test] is its flexibility in time and space. All that is required to take a WBT is a computer with a web browser and an internet connection (or the test on disk). Test takers can take the WBT whenever and wherever it is convenient, and test designers can share their test with colleagues all over the world and receive feedback.” (P. 88). Some other benefits of using computers in language testing according to Brown (1992b) and others scholars are discussed below:

Computer assisted testing are not limited in time and can be taken at the test taker’s convenient location, at convenient time, and without human intervention. Moreover, the testing procedure is less overwhelming (as compared to equivalent paper-

and-pencil tests) because the questions are presented one at a time on the screen unlike in an intimidating test booklet with hundreds of test items. As a result, many students like computers and even enjoy the testing process (Stevenson and Gross, 1991).

Many researches [Madsen, (1991),Kaya-Carton, Carton, &Dandonoli, (1991), and Laurier(1999)] have also proved that computer assisted tests require less time to finish, compared to the traditional paper and-pencil tests. Additionally, "The computer has the ability to measure time. The time which a learner takes to complete a task or even the time taken on different parts of a task, can be measured, controlled and recorded by computer." (Alderson 1990), Not only this, computer can register test taker's route through a test detailing how often s/he goes back to an assignment, how often s/he corrects his answers, when s/he asks for help etc. As a result, the teacher can comprehend the performance level of student. Another benefit of using computers in testing process is that all test takers receive precisely the same material and instructions no matter where or when they take the test.

As a result, it offers consistency and uniformity. So this uniformity helps the test takers in overcoming the fear and confusion during the test. CALT also tailors and adapts the test to the individual test taker's level of language ability by selecting the next item to which a test taker is exposed in the light of his or her response to the previous item. It allows testers to target the specific ability levels of individual students and can therefore provide more precise estimates of those abilities (Bock and Mislevy, 1982). A more accurate assessment of the test taker's language ability, with the help of psychometric calculations, is probably the most important advantage of CALT which offers infinite potentials both for teachers and learners.Computer assisted tests are shorter and require less time to finish as well as the questions submitted are neither too easy nor too difficult. It also provides improved test security. These benefits help in creating more positive attitude toward the test. Madsen's study (1986), which found that among the students taking both a paper-and-pencil test and a computer adaptive test 81% expressed a more positive attitude toward CALT, can be taken to support this. Moreover testing large number of people faster, accurately, Computers also offer test taker various helps on the screen such as the way s/he should precede, by clicking 'help' button; spelling check, help on syntactic errors in the learner's text etc. And last but not least, Computers can give immediate test results and feedback complete with a printout of basic testing statistics and accuracy in reporting test scores.

5. Problems/Constraints

Many of the problems and difficulties with computer-assisted language assessment are the same as those that plague traditional paper and pencil tests: validity, reliability, and wash back. That is, do the tests assess what they are intended to, do they consistently and reliably assign scores regardless of the test time or place, and, when the tests influence classroom practice, do they do so in a positive manner?

Other issues pertaining to CALT concern the specific testing of listening, speaking, and writing. It can be argued that listening comprehension tests can be made infinitely more authentic in a computerized environment where the incorporation of streaming video is possible; however, a common problem is the delivery of high-quality video to a large number of test takers (Buck, 2005). The necessary bandwidth for delivery can be costly, and smaller language testing programs may not have adequate resources to fund projects with the latest technology. Many multimedia players used by basic, computerized L2 listening tests allow for the test takers to play the video or audio files more than once, to rewind, or to fast forward.

There has not been much research on how this capability affects scoring or how it may alter the test construct. More sophisticated test environments can track this information; thus much research in this area is expected. Many schools and universities lack computers with audio recording equipment, and this has been a major block to the proliferation of online or computer-based speaking tests. As a result, expense is one of the major setbacks for computer administered tests of speaking ability. The major issue pertaining to CALT concerns the specific testing of writing. Writing on the computer is a cognitive process that differs greatly from the cognitive process of writing on paper, and this distinction has been a major concern among CALT researchers for some time. This issue is compounded when considering the different modes of writing (paper-and-pencil versus compute based) that are involved in writing logographic languages such as Chinese.

Issues related to adaptive item selection in CAT have also raised many concerns among scholars [Canale (1986) Carol A. According to Carol A. Chapelle and Dan Douglas(2006), "selection of items to be included on an adaptive test by an algorithm may not result in an appropriate sample of test content and may cause test takers anxiety (P.41)."CATs are based on the Item Response Theory model (IRT) which cannot be used for all item types. Therefore, CATs are not applicable to open-ended questions and items which cannot be calibrated easily (Rudner, 1998). Additionally, another important drawback of CATs is that an examinee is not allowed to go back and change answers because the next items are selected based on the previous answered items (Rudner, 1998). Moreover, item calibration is an important factor which affects the success of a CAT. If the items are not appropriately calibrated on the difficulty/ability scale, the test will be neither valid nor reliable. Additionally, at each difficulty level, several items should be used to make repeated measures at that level possible. Consequently, a large bank of items should be created (Larson, 1987). Issue related to security for high- stakes tests or identity detection of the test taker is the other negative aspect of computer assisted language testing. Another concern is related to inaccurate automatic response scoring.

To sum up, the problems/constraints of using computers in language testing result fromphysicaland performance considerations in more detail based on Brown (1992b) and other scholars are as following:

Computer equipmentmay notalwaysbe available, or be in reliable working order.

The graphics capabilities of many computers (especially older ones) may be limited.

Differences in the degree to which students are familiar withusing computers or typewriter keyboards may lead to discrepancies in their performances (Henning, 1991).

Computer anxiety is another potential problem (Henning, 1991).

6. Conclusion and Future Directions

Technological advancements have moved very rapidly since the last century. Computers have become the most useful facilitator in achieving the majority of our goals. It can be instrumental in expansion and innovation in language testing. The world of CALT will continue to develop, and this is seen "as a natural evolution in assessment practice"(Dunkel, 1999,p. 77). Testing via the computer is a logical step, in that resources are available and because computerized testing can be more motivating, streamlined, and can incorporate automatic scoring. The benefits embedded in CALT is making it integral part of today's education system to make testing practice more flexible, innovative, dynamic, efficient and individualized as well as to enhance the quality and standard of education.

Looking beyond the benefits of CALT a system of checks and balances is needed to assure that computerized tests are increasing our ability to efficiently make valid inferences about language learners' abilities and weaknesses. We must be sure computerized tests contribute overall to L2 programs and L2 learning. Computerized tests should not just increase the efficiency of test administration and scoring, but should also accurately reflect the ways in which L2s are learned and should appropriately take advantage of advances in technology to make for better testing conditions, not just different ones (Chapelle and Douglas, 2006). Also, it should be clear that the computer assisted language testing do not make a good language test without sophisticated expert knowledge of test writing. The use of technology in language testing may have its own caveats and all the negative aspects and caveats associated with CALT that mentioned so far should not lead to the suspicion towards CALT and authorities should be aware of the positive and negative aspects of a computer adaptive test when they want to administer a test in adaptive format.

The last and not the least issue in the CALT environment which will be continue for some time is scoring of essays; human raters are still needed for the scoring of extended essays, which adds a considerable expense to the otherwise monetarily efficient scoring process. However, as mentioned earlier, research has shown promise in the use of computers for rating essays, and in the future we should see computers that are able to score essays not only based on syntactic complexity, lexical complexity, and grammatical accuracy (Li, 2000), but also on discourse coherence, syntactic variety, and on-topic content. Additionally, online systems that directly or indirectly support computerized language testing, such as rater training programs, item development training sessions, and programs for uploading, revising, and finalizing items for item bank completion and maintenance, are becoming more common and should flourish in coming years. Therefore significant body of research needs to be motivated on these areas so that, in turn, the potential benefits embedded in them can be exploited for the betterment of language testing practice in general.

References

Dunkel, P. (ed.) (1991). Computer-assisted language learning and testing: Research issues and practice. New York: Newbury House.

Ockey, G. J. (2009). Developments and challenges in the use of computer-based testing for assessing second language ability. The Modern Language Journal, 93, 836–47.

Parshall, C. G., Davey, T., &Pashley, P. J. (2000). Innovative item types for computerized testing. In W. J. Van der Linden & C. A. W. Glas (Eds.), Computerized adaptive testing: Theory and practice (pp. 129–48). Dordrecht, Netherlands: Kluwer.

Wainer, H. (1990): Introduction and history. In H. Wainer (Ed.), Computerized adaptive testing: a primer (1-22). Hillsdale, NJ: Lawrence Earlbaum.

Wainer, H., &Eignor, D. (2000). Caveats, pitfalls, and unexpected consequences of implementing large-scale computerized testing. In H. Wainer (Ed.), Computerized adaptive testing: A primer (2nd ed., pp. 271–99). Mahwah, NJ: Erlbaum.

Winke, P., &Fei, F. (2008).Computer-assisted language assessment. In N. Van Deusen-

Alderson, J. C. (2000). Technology in Testing: the Present and the Future. System, 28(4),
593-603.

Baker, F. B. (1983). The basic of item response theory.Portmouth, NH: Heinemann.

Boo, J. &Vispoel, W. (2012). Computer versus paper-and-pencil assessment of educational development: A comparison of psychometric features and examinee preferences. Psychological Reports, 111, 443-460.

Douglas, D., &Hegelheimer, V. (2007).Assessing language using computer technology.Annual Review of Applied Linguistics, 27, 115–132.

Eignor, D. R. (1993).Deriving comparable scores for computer adaptive and conventional tests: An example using the SAT (Research No. 93-55). Princeton, NJ: Educational Testing Service.

Jamieson, J. M. (2005).Trends in computer-based second language assessment. Annual Review of Applied Linguistics, 25, 228–242.

Larson, J. W. & Madsen. H. S. (1985). Computer-adaptive language testing: Moving beyondcomputer-assisted testing. CALICO Journal, 2(3), 32-6.

Loyd, B. H., &Gressard, C. (1984). Reliability and factorial validity of computer attitude scales. Educational and Psychological Measurement, 44, 501-505.

Straetmans, G.J.M., &Eggen T.J.H.M. (1998, January). Computerized adaptive testing: what it is and how it works. Educational Technology, 82-89.

Young, F., Shermis, M. D., Brutten, S. & Perkins, K. (1996). From conventional to computer adaptive testing of ESL reading comprehension. System, 24(1), 32-40.

Scholl & N. H. Hornberger (Eds.), Encyclopedia of language and education (Vol. 4, pp. 353–64). New York, NY: Springer.