ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 2, No. 1, pp. 174-178, January 2012

doi:10.4304/tpls.2.1.174-178

Testing and Technology: Past, Present and Future

Salma Parhizgar
Department of Foreign Languages and Linguistics, Shiraz University, Shiraz, Iran
Email: salmaparhizgar@yahoo.com

***Abstract*—The current article is an attempt to review the history of employing technology for the purpose of assessing English as a second/foreign language. The computer-based testing's state-of-the-art will be discussed extensively in this writing. Different kinds of computer systems which were used in the past will be reviewed, the present practices will be commented on, and the future trends will be predicted.**

***Index Terms*—computer-based testing, computer adaptive testing, internet-based testing, natural language processing, automatic speech recognition**

I. Introduction

As technology mingles rapidly with every aspect of human life, it becomes more invisible in the eyes of its users. The field of language learning, teaching and testing, however, is an exception. This is where the use of technology is not yet commonplace. The recent computerized technologies used for language learning and assessment are seen as strange entities to be discovered and studied on and even to be afraid of. Therefore, the need is for educators and practitioners to become aware of the current practices in the domain of computer assisted testing.

The current article is an attempt to review the history of employing technology (mostly computers) for the purpose of assessing English as a second/foreign language. In order to reach such an aim, a summary of the commonest computer-based testing systems is provided. It is also hoped that a preliminary image of the future of computer-aided testing would be drawn at the end of this paper.

II. Computer-based Testing

The very first application of computers in testing was for scoring objective test items. In 1935, the IBM model 805 was used for the first time in the United States to reduce the costly procedures of scoring multiple-choice tests. Since the computerized scoring of the tests was thought to produce more reliable results in comparison to the previous hand-scored ones.

With the rapid change of technology during the 1970s and 1980s, language testers started to use the new computer systems for purposes beyond the simple scoring of tests. As classified by Burstein, Frase, Ginther and Grant (1996) and reported in Fulcher (2000), computers were begun to be used in eight different areas: test design, test construction, tryout, delivery, management, scoring, analysis and interpretation, and reporting. From among these usages, however, only three – delivery, scoring and reporting – were paid enough attention in the subsequent years.

In fact, the emergence of what is today known as computer-based testing (CBT) goes back to the mid-80s. Drawing on the Classical Test Theory (CTT), the first CBT tests were "simply paper-and-pencil tests delivered through the new electronic medium" (Fulcher, 2000, p.96). But it was not until the 1990s that the use of computers for development and delivery of language tests was extended. For instance, in 1998, the computer version of the Test of English as a Foreign Language (TOEFL) was presented to the US applicants. Obviously, the major negative attitude towards the computer-based version of TOEFL was the question of computer familiarity. In an attempt to find a relationship between test-takers' computer familiarity and their performance on the test, Taylor, Kirsch, Eignor and Jamieson (1999) compared examinees with high and low level of computer literacy. Their findings showed that there was no evidence of bias against candidates with low computer literacy. And as a survey by Stricker, Wilder and Rock (2004) revealed, despite examinees' differences, attitudes toward the computer-based TOEFL were moderately positive in three different countries of Argentina, Egypt, and Germany among different language learners.

Computer-based tests, in general, are considered to have some disadvantages as well as several advantages. In a comprehensive study, Alderson (2000) discusses a number of pedagogic and technical merits of CBTs. He believes that technically the computer-based tests can remove the constraints of test administration such as fixed delivery dates and locations. He also mentions that test results are available instantly and that the test security is greater as a result of examinees' access to a large database of items.

From a pedagogic point of view, again, Alderson (2000) reviews CBTs' major advantage; its user friendliness. To give an elaboration, he emphasizes that the tests would be more meaningful because of the feedback they give to the users. They also offer a range of support to learners from help facilities and dictionaries to clear instructions, examples and performance clues. In addition to these, Chalhoub-Deville (2001) also thinks that CBT allows new item/task types and tracking of student performance. According to this author, another exciting capability of CBT is the "adaptive approach"; a branch of testing which will be discussed later in this article.

Two of the main demerits of CBTs are the allowance of restricted types of items and the difficulty of assessing the highly productive skills of speaking and writing. And as mentioned before, the need for a degree of computer literacy is still of concern to many testers. A matter which led Educational Testing Service (ETS) to devise a tutorial for CBT TOEFL in an effort to remove any possible bias against computer illiterates.

III. Computer Adaptive Testing: CAT

As many researchers have described before (Stanfield, 1986; Dunkel, 1991; Fulcher, 2000; etc.), computer adaptive testing (CAT) is the most important development of the 90s. The initial CAT systems were developed in the 1970s. The driving principle underlying CAT production and use was Item Response Theory which the discussion of its distinctive models is beyond the scope of the current paper.

In any given computer adaptive test, the testers are presented with one test item at a time. If the test-taker answers the first item correctly, s/he is provided with a more difficult item. If not, an easier test item is presented to him/her. This way, the computer can adjust/*adapt* the test items to each tester's level of language ability.

Two of the most well-known CAT instruments are MicroCAT which was released in 1984, and FastTEST which was produced by Assessment Systems Corporation in 1999. Many other testing organizations also engaged in the development of CAT systems. For example, Brigham Young University developed French, German, and Spanish CAT instruments for placement at universities. As another example, The University of Cambridge Local Examinations Syndicate (UCLES) also developed CAT instruments in various languages and for various purposes.

Today CAT programs are among the most desirable testing approaches. Computer adaptive tests are in many ways advantageous over the paper-and-pencil assessments. First, they contain a large item bank which enables it to match to the testers' needs. Second, no test-taker is given the same set of test items as the other; hence, test security is increased. Third, a computer adaptive test saves time and resources as a result of reducing the number of items required to be responded. Fourth, it provides immediate results. Finally, it has the ability to distinguish learners with extremely low or high abilities (Fulcher, 2000).

In general, it could be added that:

Computer adaptive tests are often argued to be more user-friendly, in that they avoid users being presented with frustratingly difficult or easy items. They might thus be argued to be more pedagogically appropriate than fixed-format tests. (Alderson, 2000, p.596)

As Fulcher mentions in the same article in 2000, CAT has a number of disadvantages. First, providing a large number of items for the item bank is time-consuming and costly. Second, achieving totally calibrated items is not as easy as it appears to be. Third, the question remains as whether CAT is forced to include a representative sample of items in terms of content validity. Finally, unlike paper-and-pencil tests, CAT does not allow language examinees to omit items or to review them at the end of the test.

IV. Internet-based Testing

Testing on the Internet or what Roever (2001) calls web-based testing (WBT), refers to the instruments which assess language in the environment of the World Wide Web. The tests often would be downloaded on the users/clients' computer. The tests are normally the same CBTs or CATs delivered in a new medium. The kind of item types which can be presented through a WBT consists of multiple-choice items, C-tests, discourse completion tests, and reading comprehension tests accompanied with sound and video files.

But why do we need to apply WBTs while we already have enough access to a great range of computer-based assessments? Fulcher (2000), Alderson (2000), and Roever (2001) have answered this question by enumerating a number of advantages for language testing through the Internet. The main superiority of WBTs is their flexibility in time and space. Having a computer with an Internet connection, the test user can take a test whenever and wherever s/he wants.

The Web-based tests also are more flexible in terms of their design. A series of frames could represent texts, images, audio, and video at the same time. The users will have no problem in accessing to help facilities, databases, or libraries which are all updated constantly. Moreover, test results can be sent immediately to the score users. And as Roever (2001) adds, "WBTs are very inexpensive for all parties concerned" (p. 88).

The most recognizable Internet-based testing project is DIALANG funded by the European Union. The project was intended to be diagnostic at low stakes. In DIALANG, the language of administration and the skill to be tested is chosen by test users from 14 different European languages. But the program suffers greatly from limitations in testing productive skills. The most important constraint, however, is the risk of being broken by hackers; a problem which is more serious in case of the high-stakes tests of language.

As another example, the English as a Second Language Placement Examination and tests of Chinese, Japanese, and Korean at UCLA were being adapted in 2000 to be delivered on the Internet.

V. Current Research

Since the computer/Internet based tests are not providing efficient ways of assessing productive skills of speaking and writing, many researchers opt for the technologies with extensive power of natural language processing (NLP) and automatic speech recognition.

In a most recent study, Chapelle and Chung (2010) elaborated rigorously on two examples of commonly used speech rating systems – namely PhonePass and SpeechRater^{SM} v 1.0, stressing that a few attempts have been made to use the speech recognition technologies as means of testing learners' speaking abilities.

In 1997, PhonePass was investigated in detail by Bernstein. The test which is still used in some educational contexts, aimed at assessing listening and speaking abilities of second/foreign language learners. A number of tasks were included such as reading aloud, repeating sentences, and giving short answers to questions. Central to the program, was a speech recognition technology to score responses given through the telephone.

SpeechRater^{SM} is an automatic speech rating system which is now used in TOEFL Internet-based test (iBT) Speaking Practice Test. The aim is to test the users' speech as an evidence for their ability in constructing communication. Based on a multiple regression scoring model, it examines the examinees' fluency, vocabulary, grammar, and pronunciation.

Several analysts have argued for the issue of validity in relation to the SpeechRater^{SM}. For instance, Clauser, Kane, and Swanson (2002) focus on the factors to be considered in the inferences to be drawn based on test scores. And Xi (2008) takes a new approach in introducing the validation of tests as something beyond the simple correlations between scores obtained by human raters and automated procedures.

In the same vein, but this time in the area of writing assessment, e-rater was developed by Educational Testing Service (ETS) to score the examinees' essay responses. Today, e-rater is utilized as a second rater in Graduate Management Admissions Test (GMAT), in Graduate Record Examination (GRE) and in the independent writing task of the TOEFL iBT; also as the sole rater in the TOEFL Practice On-line (TPO), and Criterion.

Most of the research today, however, is focused on different aspects of the development and implementation of CAT systems. For instance, in 2009, Huang, Lin and Cheng designed an adaptive testing system which could support several assessment functions and different devices. The test could precisely measure learners' ability with large estimates of reliability and validity. Evaluation of the system's usability on the Web truly indicated that it was successful in providing an adaptive testing for different devices and supporting versatile assessment functions.

In 2009 Frey and Seitz also examined the multidimensional adaptive testing (MAT) presented by Segall (1996). They believed that "the concept of MAT is very promising for the assessment of different competencies" (p. 89). In this new approach, in addition to a great reduction in the number of items, simultaneous testing of multiple constructs was possible. According to the writers, this is "an attractive new step to more theoretically based testing that is likely to enhance the validity of test score interpretations within educational and psychological assessments" (p. 93).

In an innovative study, Lazarinis, Green and Pearson (2010) constructed a pilot study to test the capabilities of a hypermedia Web assessing tool. Based on computer adaptive testing, this framework also contains a number of "user-customizable rules" for a fully personalized assessment. These factors include learners' knowledge, educational background, goals, and preferences, as well as their performance. The advantage of this approach over the previous CAT systems is "the offered flexibility to both learners and educators. Educators are able to reflect their instructional experience in order to create tests more tailored to the characteristics of their learners. Learners can take advantage of their previous knowledge or their current goals and be examined in shorter tests with more focused items" (p. 1742).

The recent advances in the area of mobile technologies, in turn, provide suitable conditions for the use of mobile phones in the delivery of systems such as CAT. Having this in mind, Triantafillou, Georgiadou and Economides (2008) attempted to describe the design issues related to the development and implementation of a CAT on mobile devices (CAT-MD). The results showed that, according to its users, CAT-MD was an effective and efficient assessment tool; it was accurate, exact, and reliable and more importantly it was very desirable to work with since it could be used almost anywhere.

VI. Future Trends

Like all other language testing research, the concept of construct validity is central to the area of computer-based assessment. The need for reconsideration of the construct validity and generalizability of CBTs is often emphasized by the researchers in the field. Bachman (2000), for example, reviews a number of studies which examined the nature of constructs to be tested in computer assisted tests. It is always important to know exactly what the test measures. As Alderson (2000) puts it, research is needed "that will reveal more about the validity of the tests; that will enable us to estimate the effects of the test method and delivery medium" (p. 603).

In the same paper, Alderson asks for a research agenda which could help the developers to take decisions in relation to the nature of the most effective and meaningful feedback, the best ways of diagnosing strengths and weaknesses, the most appropriate clues, and the integration of media and multimedia; and also on the impact of the use of technology on learning, learners, and the curriculum.

Fulcher (2000) adds another major concern for the prospective uses of technology in language testing. According to him, the ethical aspect of computer based testing has to be considered; whether the test ranks the examinees the same way as paper-and-pencil tests do, and whether factors such as age, gender, educational background, previous experience, and attitudes to technology affect the test scores or not.

Throughout the same article, Fulcher reminds us of a predicted fourth generation of assessment known to calculate trajectories in language learning. Taking advantage of the advances in artificial intelligence (AT), the system would be able to predict the progress of individuals in a meaningful way. Alderson (2000) along with Chalhoub-Deville (2001), also, hope that in the near future many of today's open-ended productive tasks could be tested meaningfully by means of computers.

Referring to some of the shortcoming of NLP and speech recognition technologies, Xi (2010) concludes that computers can improve the effectiveness of language assessments, only if they will be used appropriately and responsibly. Last but not least, Roever (2001) stresses that "the Web greatly expands the availability of computer-based testing with all its advantages and will undoubtedly become a major medium of test delivery in the future" (p. 92).

VII. Conclusion

The computer-based testing's state-of-the-art was discussed extensively in this writing. Different kinds of computer systems which were used in the past were reviewed, the present practices were commented on, and the future trends were predicted.

It is worth noticing that the innovation and flexibility present in the CBTs, CATs, and other computer-based assessing systems should not result in the ignorance of the problems associated with these new mediums. All the users and developers are needed to become knowledgeable and comfortable with the new test administrations. The researchers also have the responsibility to inform others about the different aspects of the newly introduced testing procedures. On the whole, it is obvious that implementation of computer-based testing systems require a great amount of research and expertise. Therefore, without enough degree of proficiency in the area, it is far better for test administrators and examiners to remain with the same familiar conventional tests.

References

[1] Alderson, J. C. (2000). Technology in testing: The present and the future. *System* 28, 593-603.
[2] Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language testing* 17.1, 1-42.
[3] Bernstein, J. (1997). Scoring rubrics: Using linguistic description to automatically score free responses. In A. Huhta, V. Kohonen, L. Lurki-Suonio & S. Louma (eds.), *Current developments and alternatives in language assessment.* Finland: Jyvaskyla University.
[4] Burstein, J., L. T. Frase, A. Ginther & L. Grant. (1996). Technologies for language assessment. *Annual Review of Applied Linguistics*16, 240-260.
[5] Chalhoub-Deville, M. (2001). Language testing and technology: Past and future. *Language Learning and Technology* 5.2, 95-98.
[6] Chapelle, C. A. & Y. R. Chung. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing* 27.3, 301-315.
[7] Clauser, B. E., M. T. Kane & D. B. Swanson. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education* 15, 413–432.
[8] Dunkel, P. (ed.) (1991). Computer assisted language learning and testing: Research issues and practice. New York: Newbury House.
[9] Frey, A. & N. N. Seitz. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation* 35, 89-94.
[10] Fulcher, G. (2000). Computers in language testing. In P. Brett & G. Moterram (eds.), *A special interest in computers: Learning and teaching with information and communications technologies*. Manchester: IATEFL Publications, 93-107.
[11] Huang, Y. M., Y. T. Lin & S. C. Cheng. (2009). An adaptive testing system for supporting versatile educational assessment. *Computers and Education* 52, 53-67.
[12] Lazarinis, F., S. Green & E. Pearson. (2010). Creating personalized assessments based on learner knowledge and objectives in a hypermedia Web testing application. *Computers and Education* 55, 1732-1743.
[13] Roever, C. (2001). Web-based language testing. *Language Learning and Technology* 5.2, 84-94.
[14] Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika* 61, 331–354.
[15] Stansfield, C. (1986). Technology and Language Testing. Washington, DC: TESOL.
[16] Stricker, L. J., G. Z. Wilder & D. A. Rock. (2004). Attitudes about the computer-based test of English as a foreign language. *Computers in Human Behavior* 20, 37-54.
[17] 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, 219-274.
[18] Triantafillou, E., E. Georgiadou & A. A. Economides. (2008). The design and evaluation of a computerized adaptive test on mobile devices. *Computers and Education* 50, 1319-1330.
[19] Xi, X. (2008). What and how much evidence do we need? Critical considerations for using automated speech scoring systems. In C. A. Chapelle, Y. R. Chung & J. Xu (eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment*. Ames, IA: Iowa State University, 102-114.
[20] Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing* 27.3, 291-300.

Salma Parhizgar (b. 1983, Shiraz, Iran) is an M.A. candidate of TEFL at Shiraz University, International Branch. In 2009, she received her B.A. in English Translation from the Islamic Azad University of Shiraz. Her research interests include Computer Assisted Language Learning, testing and assessment, contrastive analysis, and pragmatics.