INTERNATIONAL JOURNAL OF HIGHER EDUCATION AND RESEARCH
IJHER, Vol., 9(1), 2019, 1-11.
www.ijher.com

IJHER
(ISSN 2277 260X)

CODEN:IJHER ORIGINAL ARTICLE

COMPUTER ASSISTED LANGUAGE TESTING (CALT): ISSUES AND CHALLENGES

Zain Mohammad Sulaiman[1] and Mohd Ahmad Khan[2]
[1](Department of Linguistics, Aligarh Muslim University, India, zainsulaiman89@gmail.com)
[2](English Lecturer, Syscoms College, Abu Dhabi, UAE, ahmadsyscoms@gmail.com)

ABSTRACT

In today's world information and communication technology has almost integrated with every human work. Computer along with the internet have become one of the most important tool and has sparked a revolution and made the current era a digital age. The integration of computer to language learning and teaching is the most commonly practiced mode of language education worldwide. Computer-assisted language testing (CALT) employs computer applications eliciting and evaluating test takers' performance in a second language. CALT encompasses computer-adaptive testing (CAT), the use of multimedia in language test tasks, and automatic response analysis (Chapelle & Douglas, 2006). While learning and teaching a language, especially a foreign language, becomes the most essential part. The three main motives for using technology in language testing are efficiency, equivalence, and innovation. The paper aims to highlight the detailed description of CALT along with its various dimensions; its application and the methods involved. It will also throw lights on assessing English for Specific Purposes (ESP). Since technology is a challenging task, particularly computer and language teaching and learning, it will explore the challenges of CALT, referring to Indian Classrooms.

Keywords: CALT, Learning, Teaching, Issues and Challenges

INTRODUCTION TO CALT

José Noijons (1994) defines CALT is an integrated procedure in which language in which language performance is elicited and assessed with the help of a computer. CALT encompasses computer-adaptive testing (CAT), the use of multimedia in language test tasks, and automatic response analysis (Chapelle & Douglas, 2006). Chapelle (2010) distinguishes three main motives for using technology in language testing: efficiency, equivalence, and innovation.

José Noijons (1994) defines CALT is an integrated procedure in which language in which language performance is elicited and assessed with the help of a computer. CALT encompasses computer-adaptive testing (CAT), the use of multimedia in language test tasks, and automatic response analysis (Chapelle & Douglas, 2006). 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).

Table1.1 Framework for Computer Assisted Language Tests

Attribute	Categories
Directionality	Linear, adaptive, and semi-adaptive testing
Delivery format	Computer-based and Web-based testing
Media density	Single medium and multimedia
Target skill	Single language skill and integrated skills
Scoring mechanism	Human-based, exact answer matching, and analysis-based scoring
Stakes	Low stakes, medium stakes, and high stakes
Purpose	Curriculum-related (achievement, admission, diagnosis, placement, progress) and non-curriculum-related (proficiency and screening)
Response type	Selected response and constructed response
Task type	Selective (e.g., multiple choice), productive (e.g., short answer, cloze task, written and oral narratives), and interactive (e.g., matching, drag and drop)

CALT: ORIGIN AND DEVELOPMENT

The use of computer in the field of assessment and testing practice dates back to 1935 when the IBM model 805 was used for scoring objective tests in the United States of America to reduce the labour intensive and costly business of scoring millions of tests taken each year. But the year 1980is a crucial year which led to many advancements in the area of CALT. In the 1980s,

as the microcomputers came within reach for many applied linguists and item response theory (IRT) also appeared at the same time to make use of this new technology for innovating the existing assessment and testing practice. In 1985, Larson and Madsen developed the first CAT at Brigham Young University, in the USA which was technologically advanced assessment measures (Dunkel, 1999). They developed large pool of test items for test delivery using computers. In the Computer Adapted Test, designed by them, the program selected and presented items in a sequence based on the test taker's response to each item. If a student answered an item correctly, a more difficult item was presented; and conversely, if an item was answered incorrectly, an easier item was given. In short, the test "adapted" to the examinee's level of ability. The computer's role was to evaluate the student's response, select an appropriate succeeding item and display it on the screen. The computer also notified the examinee of the end of the test and of his or her level of performance (Larson 1989: 278). Larson and Madsen's (1985) above referred CAT served as an impetus for the construction and development of many more computer adapted tests throughout the 1990s (e.g., Kaya-Carton, Carton & Dandonoli, 1991; Burston & Monville-Burston, 1995; Brown & Iwashita, 1996) which helped language teachers in making more accurate assessment of the test taker's language ability and attracted many as it appeared to be of immense potentials both for language teachers and learners. As Item Response Theory and many computer softwares, for calculating the item statistics and providing adaptive control of item selection, presentation and evaluation, witnessed advancements, the use of computer technology in the field of language assessment and testing started becoming inevitable reality though the challenge of availability of infrastructure and the cross-disciplinary knowledge, required in the field, hampered its progress for some time at its early stage.

Today the use of computer technology, in the field of language assessment and testing, has become so widespread and so inclusive that it is regarded as the inseparable part of today's education system. The web of many useful computer adapted tests [CATs] as well as web based tests [WBTs] is constantly growing and computers are used not only for test delivery but also for evaluation of complex types of test responses. Even the large testing companies, who showed little interest in the field at its early stage, have also stepped in and are producing and

administrating these CATs as well as WBTs. The administration and delivery of highly popular and useful tests such as TOEFL, IELTS, DIALANG etc., to mention a few, speak volumes about the role played by computer technology in the field of language assessment today.

PROMINENT TESTING SERVICES

The realm of CALT is constantly expanding and encompassing even the field of scoring and rating as well. Today computers are used not just to score objective type of test tasks but also to assess and rate much more complex task types like essays and spoken English. The Educational Testing Service's http://www.ets.org automated systems known as Criterion http://www.criterion.ets.org and e-rater http://www.ets.org/erater for rating extended written responses based on aspects of NLP analysis, Vantage Laboratories' http://www.vantage.com IntelliMetric, Pearson Knowledge Technologies' http://www.knowledge-technologies.com Intelligent Essay Assessor (IEA), and Pearson's Versant, http://www.versanttest.com a computer-scored test of spoken English for non-native speakers, using NLP technology, etc. indicate how rapidly the realm of CALT is growing and reshaping, innovating and revolutionizing the field of language assessment and testing by adapting itself successfully with the new challenges in technology and assessment practice .

EVALUATION IN CALT

Testing and Evaluation is the most important part of language learning because without learning process there can be test. The systematic evaluation can be done by recognizing the influence on learning of three main perspectives (software designer, teacher and student) and taking into account three sets of interactions between them:

- Teacher-student: a two-way direct interaction.

  One of the main variables here is the teacher's role, which may be 'resource provider', 'manager', 'coach',' researcher' or 'facilitator'.
- Designer-student

  Primarily a one-way influence, although the designer's perception of the student's learning characteristics will implicitly be of help.
- Designer-teacher

Again, primarily a one-way influence, with the designer's perception of the teacher having some influence. This framework assists the evaluator to identify the key issues on which judgements must be made in the particular context of the proposed use (predictive evaluation) or actual use (interpretive evaluation). (Soromic, 2010).

CALT IN ESP CLASSROOM

The application of technology in the realm of English for Specific Purposes (ESP) has gained tremendous popularity among English as a Foreign Language (EFL) researchers and scholars (Arno, 2012; Butler-Pascoe, 2009). ESP instruction is goal-oriented and based on the specific needs of students (Robinson, 2003).

Corpus helps to test the communicative ability and efficiency. Content, language, grammar and vocabulary knowledge is being assessed. The assessment of curriculum, instructional materials are constantly assessed. The most important part of testing involves the language usage for a specific purposes, i.e. business, medical, law, science and technology, etc. and the usage of vocabulary. The assessment of curriculum development is the primary task.

CHALLENGES IN CALT

The views regarding the current status and the future of CALT vary slightly among researchers, with some being more concerned about the severity of existing problems than others. Ockey (2009), for instance, believes that due to numerous limitations and problems "CBT has failed to realize its anticipated potential" (p. 836), while Chalhoub-Deville (2010) contends that "L2 CBTs, as currently conceived, fall short in providing any radical transformation of assessment practices" (p. 522). In the meantime, other researchers (e.g., Chapelle, 2010; Douglas, 2010) appear to be somewhat more positive about the transformative role of CALT and stress that despite existing unresolved issues technology remains "an inescapable aspect of modern language testing" and its use in language assessment "really isn't an issue we can reasonably reject—technology is being used and will continue to be used" (Douglas, 2010, p.139). Still, everyone seems to acknowledge the existence of challenges in

CALT, maintaining that more work is necessary to solve the persisting problems. In particular, a noticeable amount of discussion in the literature has been dedicated to the issues plaguing computer-adaptive testing, which, according to some researchers, led to the decline of its popularity, especially in large scale assessment. Of primary concern for CATs is the security of test items (Wainer & Eignor, 2000). Unlike a linear CBT that presents the same set of tasks to a group of test takers, a computer adaptive language test provides different questions to test takers. To limit the exposure of items, CATs require a significantly larger item pool, which makes the construction of such tests more costly and time-consuming. Ockey (2009) suggests that one way to avoid problems associated with test takers' memorization of test items is to create computer programs that would generate questions automatically. Some test developers suggest starting a CAT with easy items, whereas others recommend beginning with items of average difficulty. Additionally, no consensus has been reached on how the algorithm should proceed with the selection of items once a test taker has responded to the first question, nor are there agreed-upon rules on when exactly an adaptive test should stop (Thissen & Mislevy, 2000). Nonetheless, research is being carried out to address this issue and new methods of item selections in computer-adaptive testing such as the Weighted Penalty Model (see Shin, Chien, Way, & Swanson, 2009) have recently been proposed. Another major problem with computer-adaptive tests concerns their reductionist approach to the measured L2 constructs. Canale (1986) was one of the first to argue that the unidimensionality assumption deriving from the IRT models used in CATs poses a threat to the L2 ability construct, making it unidimensional as well. Their main argument suggests that the L2 ability construct should be multidimensional and consist of multiple constituents that represent not only the cognitive aspects of language use, but also knowledge of language discourse and the norms of social interaction, the ability to use language in context, the ability to use metacognitive strategies, and, in the case of CALT, the ability to use technology. Hence, Chalhoub-Deville (2010) asserts that, because of the multidimensional nature of the L2 ability construct, measurement models employed in CBTs must be multidimensional as well—a requirement that many adaptive language tests do not meet. Finally, the unidimensionality assumption of IRT also precludes the use of integrated language tasks in computer-adaptive assessment (Jamieson, 2005).

As a result of some of these problems, ETS, for instance, decided to abandon the computer-adaptive mode that was employed in TOEFL CBT and instead return to the linear approach in the newer TOEFL iBT. The limitations of the adaptive approach prompted some researchers to move toward semi adaptive assessment (e.g., Winke, 2006). The advantages of this type of assessment include a smaller number of items (compared to linear tests) and the absence of necessity to satisfy IRT assumptions. Thus, Ockey (2009) argues that semi adaptive tests can be the best compromise between adaptive and linear approaches and predicts that they will become more widespread in medium-scale assessments.

Automated scoring is another contentious area of CALT. One of the main issues with automated scoring of constructed responses, both for writing and for speaking assessment, is related to the fact that computers look only at a limited range of features in test takers' output. Even though research studies report relatively high correlation indices between the scores assigned by AWE systems and human raters (e.g., Attali & Burstein, 2006), Douglas (2010) points out that it is not clear whether the underlying basis for these scores is the same. Specifically, he asks, "are humans and computers giving the same score to an essay but for different reasons, and if so, how does it affect our interpretations of the scores?" (Douglas, 2010, p. 119). He thus concludes that although "techniques of computer-assisted natural language processing become more and more sophisticated, . . . we are still some years, perhaps decades, away from being able to rely wholly on such systems in language assessment" (Douglas, 2010, p. 119). Since machines do not understand ideas and concepts and are not able to evaluate the meaningful writing, critics contend that AWE "dehumanizes the writing situation, discounts the complexity of written communication" (Ziegler, 2006, p. 139) and "strikes a death blow to the understanding of writing and composing as a meaning-making activity" (Ericsson, 2006, p. 37).

Automatic scoring of speaking skills is even more problematic than that of writing. In particular, speaking assessment involves an extra step which writing

assessment does not have: recognition of the input (i.e., speech). Unlike writing assessment, the assessment of speaking also requires the evaluation of segmental features (e.g., individual sounds and phonemes) and suprasegmental features (e.g., tone, stress, and prosody). Since automated evaluation systems cannot perform at the level of human raters and cannot evaluate coherence, content, and logic the way humans do. Other challenges faced by CALT are related to task types and design, namely the use of multimedia and integrated tasks. Although the use of multimedia input is believed to result in a greater level of authenticity in test tasks by providing more realistic content and contextualization cues, it remains unclear how the inclusion of multimedia affects the L2 construct being measured by CBTs (Jamieson, 2005). Some researchers even question the extent to which multimedia enhances the authenticity of tests (e.g., Douglas & Hegelheimer, 2007) since comparative studies on the role of multimedia in language assessment have yielded mixed results (Ginther, 2002; Wagner, 2007; Suvorov, 2009). With regards to integrated tasks, their implementation in CBTs is generally viewed favorably because such tasks seem to better reflect what test takers would be required to do in real-life situations. The use of integrated tasks is therefore believed to increase authenticity of language tests (Fulcher & Davidson, 2007). However, Douglas (2010) warns that the interpretation of integrated tasks can be problematic because, if the test taker's performance is inadequate, it is virtually impossible to find out whether such performance is caused by one of the target skills or their combination. This concern appears to be more relevant in high stakes testing than in low stakes testing.

CONCLUSION

To sum up, all the negative aspects and caveats associated with CALT mentioned so far are worthy of concern and research but they should not lead to the suspicion towards CALT. Technology can be instrumental in expansion and innovation in language testing. Since its advent, CALT has changed and innovated the existing testing practices, to make them in line with the needs of the 21st century e-generation of second language learners by making them

more flexible, innovative, individualized, efficient and fast. The realization of these benefits embedded in it and their implications, 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. In the form of CALT, we are witnessing these opportunities for the reflections and need to capitalize on.

REFERENCES

Alderson, J. C. (1988). Innovations in language testing: Can the microcomputer help? Special Report No 1 Language Testing Update. Lancaster, UK: University of Lancaster.

Alderson, J. C. (1990). Learner-centered testing through computers: Institutional issues in individual assessment in J. de Jong & D. K. Stevenson (eds.) Individualizing the assessment of language abilities. Clevedon, UK: Multilingual Matters.

Alderson, J. C. (2000). Assessing reading. Cambridge: Cambridge University Press

Arno, E. (2012). The role of technology in teaching languages for specific purposes courses. Modern Language Journal, 95, 88–103.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v. 2. The Journal of Technology, Learning, and Assessment, 4(3), 2–29.

Burston, J. & Monville-Burston, M. (1995). Practical design and implementation considerations of a computer-adaptive foreign language test: The Monash/ Melbourne French CAT. CALICO Journal, 13(1), 26-46

Brown, A. &Iwashita, N. (1996).The role of language background in the validation of a computer- adaptive test. System, 24(2), 199-206.

Bulter-Pascoe, M.E. (2009). English for specific purposes (ESP), innovation, and technology. English education and ESP, 1-15.

Carol A. Chapelle and Dan Douglas (2006) Assessing language through computer technology Cambridge: Cambridge University Press. Center for Applied Linguistics: www.cal.org accessed on 10 June, 2012.

Chalhoub-Deville, M. & Deville, C. (2010). Computer adaptive testing in second language contexts. Annual Review of Applied Linguistics, 19, 273-99.

Canale, M. (1986). The promise and threat of computerized adaptive assessment of reading comprehension. In C. Stansfield (ed.), Technology and language testing (pp. 30-45). Washington, DC: TESOL

Douglas, D. (2010). Understanding language testing. London, England: Hodder Education.

Dan Douglas and Volker Hegelheimer, (2007). Annual Review of Applied Linguistics 27, 115–132.

Dunkel, P (1999). Research and development of a computer-adaptive test of listening comprehension in the less commonly-taught language Hausa. In M. Chalhoub-Deville (ed.), Development and research in computer adaptive language testing (pp.91-121) Cambridge: University of Cambridge Examinations Syndicate/Cambridge University Press.

Ericsson, P. (2006). The meaning of meaning: Is a paragraph more than an equation? In P. Ericsson & R. Haswell (Eds.), Machine scoring of student essays: Truth and consequences (pp. 28–38). Logan: Utah State University Press.

Fulcher, G. and Davidson, F. (2007). Language Testing and Assessment: An Advanced Resource Book. London and New York: Routledge.

Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. Language Testing, 19(2), 133–167.

Howell, S.L. and Hricko, M. (eds.): 2006, Online Assessment and Measurement: Case Studies from Higher Education, K-12 and Corporate. Idea Group, Hershey, PA

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

Kaya-Carton, E., Carton, A. S. & Dandonoli, P (1991). Developing a computer- adaptive test of French reading proficiency. In P. Dunkel (ed.), Computer- assisted language learning and testing: Research issues and practice (pp. 259-84) New York: Newbury House.

Larson, Jerry W.; Madsen, Harold S. (1985) CALICO Journal, 2(3), pp 32-36

Noijons, J. (1994). Testing computer assisted language tests: Towards a checklist for CALT. CALICO Journal, 12(1), 37-58.

Ockey, G.J. (2009). Developments and Challenges in the Use of Computer-Based Testing for Assessing Second Language Ability. Modern Language Journal, 93

Pearson: www.market-leader.net, www.ecollege.com, www.myenglishlab.com accessed on 15 June, 2012.

Robinson, P. (1991). ESP today. New York: Prentice Hall.

Reid, J. (1986). Using the Writer's Workbench in composition teaching and testing. In C. Stansfield (ed.), Technology and language testing (pp. 167—88). Washington, DC: TESOL Publications.

Soromic, K. (2010). Software Design: Full Consideration of Pedagogical Involvement. Newsletter Hofy.

Suvorov, R. (2009). Context visuals in L2 listening tests: The effects of photographs and video vs. audio-only format. In C. A. Chapelle, H. G. Jun, & I. Katz (Eds.), Developing and evaluating language learning materials (pp. 53–68). Ames: Iowa State University.

Shin, C., Chien, Y., Way, W. D., & Swanson, L. (2009). Weighted penalty model for content balancing in CATS. Pearson.

Thissen & Mislevy, (2000). Testing algorithm. In computerized adaptive testing: A premier, Ed.

Wainer, H. & Eignor, D). (2000). Caveats, pitfalls and unexpected consequences of implementing large-scale computerized testing. In H. Wainer etal. (2000). Computer adaptive testing: A primer 2nd edn. (pp. 271-99). Hillsdale, NJ: Lawrence Erlbaum Associates.

Winke, P., (2006). Review of hot potatoes. Language Learning & Technology, 5(2), 28–33

Wagner, E. (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. Language Learning & Technology, 11(1), 67–86.

Ziegler, W. (2006). Computerized writing assessment: Community college faculty fine reasons to say "not yet." In P. Ericsson & R. Haswell (Eds.), Machine scoring of student essays: Truth and consequences (pp. 138–46). Logan: Utah State University Press.

WEB-LINKS

- http://www.ets.org
- http://www.criterion.ets.org
- http://www.ets.org/erater
- http://www.vantage.com
- http://www.knowledge-technologies.com
- http://www.versanttest.com