European Journal of Education Studies
ISSN: 2501 - 1111
ISSN-L: 2501 - 1111
Available on-line at: www.oapub.org/edu

doi: 10.5281/zenodo.576047 Volume 3 | Issue 5 | 2017

COMPUTER ADAPTIVE TESTING (CAT) DESIGN; TESTING ALGORITHM AND ADMINISTRATION MODE INVESTIGATION

Hooshang Khoshsima[1i]
Seyyed Morteza Hashemi Toroujeni[2]
[1]Associate Professor, English Language Department,
Chabahar Maritime University, Chabahar, Iran
[2]M.A. in TEFL, English Language Department
Chabahar Maritime University, Chabahar, Iran

Abstract:
Since the advent of technology to transform education, the use of computer technology has pervaded many areas of fields of study such as language learning and testing. Chapelle (2010) distinguishes three main motives for using technology in language testing: efficiency, equivalence and innovation. Computer as a frequently used technological tool has been widely inspected in the field of language assessment and testing. Compute-adaptive language test (CALT) is a subtype and subtest of computer-assisted language test because it is administered at computer terminal or on personal computer. The issue that currently needs more attention and prompt investigation of researchers is to study the testing mode and paradigm effects on comparability and equivalency of the data obtained from two modes of presentation, i.e. traditional paper-and-pencil (PPT) and computerized tests. To establish comparability and equivalency of computerized test with its paper-and-pencil counterpart is of importance and critical. Then, in this study, the researcher indicate that in order to replace computer-adaptive test with conventional paper-and-pencil one, we need to prove that these two versions of test are comparable, in other words the validity and reliability of computerized counterpart are not violated.

Keywords: computer adaptive testing (CAT), computer adaptive language testing (CALT), testing mode administration, testing paradigm

[i] Correspondence: email khoshsima2002@yahoo.com, hashemi.seyyedmorteza@gmail.com

1. Introduction

Technology is not just a background to our lives: rather it is a major factor affecting all aspects of individual and communal lives; it is an omnipresent feature of our lives that shapes the way we work, play, live, think, and how we interact with each other (Kranzberg and Davenport, 1972). In such a way, it has a tremendous influence on our daily life and consequently brings about life-long changes at an ever increasing rate. Technology has changed the way we live our lives and is still reshaping our lives constantly and influencing every aspect of our lives (Philbin, 2003, Challoner, 2009). According to the assessment researcher, Stuart Bennett- a quite committed and enthusiastic proponent of technology- who is interested in doing research in measurement writes about the transformative influence of technology on large-scale instructional and educational assessment. He declares that serious changes in assessment are possible by the technology. The technology and technological tools permit us to create tests based on the conceptualizations of requirements and needs to make someone able to succeed in a special domain, to make practical assessment of test-takers performance through the use of computer-based simulation, automatic item generation, and automated essay scoring, and they also make it possible to transform the purposes for which we use high-stakes tests and test delivery methods (Bennett, 1999, p. 11). The seeds for idea of the transformative potential and capability of technology for assessment domain -mentioned by Bennett- have been planted and hinted at much earlier by researchers in educational measurement for years (e.g., Bejar & Braun, 1994).

The IBM model 805 machine used in 1935 has been recorded as the first attempt to use computers in testing domain. It aimed to score objective tests of millions of American test takers each year. Use of computer in language testing has resulted in the birth of independent discipline named CALT (Computer-Assisted Language Testing) which has been accelerated by CALL (Computer-Assisted Language Learning). CALT has changed the nature of language assessment field with its potential benefits and advantages. According to Jose Noijons (1994), CALT is "*an integrated procedure in which language performance is elicited and assessed with the help of a computer*" (P.38). First application of computer and the related technologies in language assessment and testing field dates back to 1953 when the objective tests in the USA were scored by the IBM model 805 in order to ease the scoring difficulties and to incorporate the new-type tests in the assessment by scoring the multiple choice items. CALT is reshaping and restructuring the very nature of language assessment not only by highly

individualizing the assessment process but also by helping overcome many of the administrative and logistical problems prevailing in the field of conventional testing (Pathan, 2012).

1.1 Computer Adaptive Language Testing (CALT)

Jose Noijons (1994) defined CALT as: An integrated procedure in which language performance is elicited and assessed with the help of a computer (P.38). Due to the nature of technology, his kind of computer use as the most prevalent technological tool in educational contexts to elicit and assess language performance in testing domain is categorized under three subfields including 1) use of computer for generating tests automatically, 2) interaction of testee with computer, and 3) use of computer for the evaluation of test taker's responses. CALT is the computer-adaptive subtype of computer-assisted tests that has three additional characteristics: (a) the test 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 (Madsen, 1991; Wainer, 1990). This flexi-level strategy eliminates the need (usually present in traditional fixed-length paper-and-pencil tests) for students to answer numerous questions that are too difficult or too easy for them. In fact, in a CALT, all students take tests that are suitable to their own particular ability levels-tests that may be very different for each student (Larson and Madsen, 1985).

1.2 Testing Mode

It is defined as the media kind through which the test is delivered to test takers. Different modes of test administration include: (1) traditional paper and pencil administration method, handled through postal services in absence of any administrator, or in presence of test administrators in a place where test is implemented (2) computer-assisted (electronic) method by use of a computerized interface (Manip Ther, 2010).

1.3 Testing Paradigm

Testing paradigm includes linear and adaptive kinds of testing strategies. Way (2010) used the term "Barely Adaptive Test" to refer to a partially targeted test used as an interim step when transitioning to CAT. In a CAT program, test items are selected based on the relative ability of the test takers according to their correct or incorrect answers given to the items (e.g., high vs. low ability), but they are not precisely targeted

to the exact ability estimate. Using the randomized item selection procedures discussed by Kingsbury & Zara (1989) and Bergstrom, Lunz, & Gershon (1992), it is possible to widen the item selection criteria to include relatively large numbers of items. As Muckle et al (2008) discussed, this process can dramatically improve pool use while minimally impacting test precision. Linear tests are similar to paper test forms in that the same set of test items is administered to all test takers who receive a given test form. CBT also typically administers only 1 item at a time. In both paper and CBT administration of linear forms, a limited number of parallel forms containing non-overlapping or partially overlapping item sets are typically constructed (Becker & Bergstrom, 2013).

1.4 Test Equivalency

Translations of paper and pencil assessment tools into computerized versions often require that the computerized form be comparable to the conventional paper and pencil one and the scores and the results obtained from two identical test forms approximate to each other. In fact, the validity of a computer based version of a test must be proved by the same methods of validity determination for its traditional counterpart that pointed out by 1985 Standards of Guidelines. Since computerized forms of standardized tests are making available, users may have the choice between taking the test in either mode.

AERA (American Educational Research Association) asserts that the equivalency of test scores from two administrations using different testing modes cannot be easily taken into granted. The equivalence between onscreen and paper and pencil tests is necessary but to show and prove the presence of equivalence between computerized and paper and pencil versions of the same test is the responsibility of test developers (Bugbee, 1996). It is important to establish score and construct equivalency if the scores and results of paper and pencil and computerized tests are to be interchangeable (MacDonald, 2002).

A research conducted by Ward, Hooper and Hannafin (1989) showed that although there was a significant difference in anxiety level with those who took the computerized version of test who had higher anxiety level, no difference in test performance between computer and paper-and-pencil based testing has been seen. The subjects participated in the study were 50 university students who were majoring in an Advanced Special Education course. 75% of the group who took the computerized testing either firmly agreed or agreed that testing via computer medium was more difficult than paper-and-pencil or traditional methods, but actually computer based

testing did not lead the test takers to any difficulty and deficiency which they were unable to overcome (p.331).

Some studies examined the score equivalence of computerized version of a reading test and conventional one. One of these studies is the research done by Mark Pomplun, Sharon Frey, and Douglas F. Becker. Pumplun, Frey & Becker (2002) studied the score equivalence of currently used paper-and-pencil version of a Nelson-Denny speeded test of reading comprehension and a new computerized version. In computer based test, the same sets of test items and questions as the paper and pencil based version were delivered to the test takers in the same order and format. And unlike the adaptive testing in which each examinee takes a set of items whose difficulty is appropriate to their estimated ability level, test items in this linear paradigm of testing were given to examinees individually on the computer screen and all the students received the same items. In the study done by Pomplun, Prey, & Becker, differences in mean scores, variances, raw score correlations and correlations corrected for unreliability were examined.

2. Literature Review

A form of computer based test with more sophisticated estimating algorithms called Computer Adaptive Language Testing (CALT) attracted much attention from language test developers during the last decade. In CALT, in order to maximize the accuracy and carefulness of the test based on known information recognized by analysing previous answered questions, new questions are selected and presented to the test takers to be answered (Hashemi Toroujeni, 2016; Weiss, D. J., & Kingsbury, G. G. 1984). This type of testing that adapts to the examinee's ability level by tailoring the difficulty of the exam has been the focus of much research in the field of psychometrics in recent years.

The reasons of failing to look after this type of testing for some decades are basically two fold. One reason has to do with the special testing paradigm applied to design, analysis, and scoring of tests called Item Response Theory (IRT) and its application to language testing. In fact, the application of Item Response Theory to language testing had been delayed by the controversial issue of language ability dimensionality (Bachman, 1990; Bachman et al., 1995; Canale, 1983; Choi, 1991; Cummins, 1983; Hashemi Toroujeni, 2016). The second significant reason to delay the application of IRT testing strategy to language testing domain is related to many issues including those that are concerned to the influences of multimedia on interactiveness

(Bachman and Palmer, 1996), the effects that will influence the performance of test takers in language testing context.

Despite the increasingly growing interest in CALT, very few studies have been done on the comparability between paper and pencil testing and computer adaptive language testing. As advantages of CALT over conventional paper and pencil testing, testing efficiency and accuracy in identifying cognitive ability or ability limits can be mentioned (Mason, Patry, Bernstein, 2001). In addition to the aforementioned benefits, the other advantages of computer adaptive language testing include: the requirement of fewer test items to arrive at a more accurate estimate of test takers' language proficiency, finer distinctions than total number correct due to the unique scoring system of CALT, higher security of test due to administering a different set of test items for each students, availability of immediate feedback, reduction of scheduling and supervision concerns for high stakes tests due to the possibility of individual administration, improvement of test taking motivation, reduction of average test score differences across ethnic groups (Pine, Church, Gialluca, and Weiss, 1979; Pine and Weiss, 1978) and storage of test takers' performance data to be tracked over time. Since test items that are above the current ability and proficiency level test takers are prohibited to be administered, the time required to finish a CALT test is shorter than its conventional counterpart.

One comparative study showed that CALT required only one-fourth of the testing time required by the paper administered tests, while the computer administered tests required only one-half to three-quarters of the testing time required by the paper administered tests (Olsen, Maynes, & Slawson, 1989). It is concluded that both types of computerized testing including linear and adaptive ones compared to the conventional version of test lead to the reduction of testing time, but CALT results in greater reduction is more considerable. Higher reliability with fewer items and higher effective life span of test items can be mentioned as other advantages of CALT. Nevertheless, researches show that CALT is not necessarily ideally suited to all types of testing such as high stakes testing. And not allowing the test takers to return to unanswered questions and to review and revise answered ones seems to be the main drawback of CALT.

Conventional testing provides poor measurement because the items have little relevance for the test takers with different ability levels who take the test. Time limit imposed on traditional tests is another drawback of this type of test that often worsens the quality of assessment.

It seems that the recognized fundamental measurement problems that characterize traditional paper and pencil tests have been resolved by introduction of Computer Adaptive Language Testing (CALT). The principle of adapting the test items to the test takers applied in the CALT was used for the first time by the developer of Binet IQ test, Alfred Binet (Binet & Simon, 1905). Although the U.S. Army's researchers' first attempts to construct elementary and undeveloped adaptive tests using both paper and pencil and testing machines (Bayroff, 1964; Bayroff, Thomas, & Anderson, 1960) turned unsuccessful, the Binet IQ test remained the only operational adaptive test for more than a half of century. To advocate theoretical research on adaptive testing and item response theory, another attempt was done by Fredric Lord (e.g., Lord, 1970, 1971a) in the late 1960.

Larson and Madsen are the ones who developed the first CAT project at Brigham Young University, in the USA. They designed the first Computer Adaptive Test by developing a large pool of test items with the help of computer for test delivery in which the program adapted and presented items in a sequence and order based on the test taker's response to each item. Based on Item Response Theory (IRT) paradigm applied in CAT program to design, analyze and score the test, the CAT testing program will adapt and present a more difficult item if the tester answers an item correctly and conversely, an easier one will be selected and presented to the tester if the test item is answered incorrectly. This happens based on the tester's level of ability and knowledge. 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 & Madson, 1989).

The International Guidelines on Computer-Based Testing (International Test Commission 2004) stated that to establish a valid and reliable computerized test and replace it with its paper-and-pencil counterpart, equivalent test scores of two versions should be established. It is exactly what the comparability of CALT and PPT means.

This set of testing standards is supported by the classical true-score test theory – the basis of computer-based and paper-based testing (Allen & Yen 1979). Under this theory, a test taker who takes the same test in the two modes is expected to obtain nearly identical test scores. The standards are also supported by empirical studies (Khoshsima, Hosseini & Hashemi Toroujeni, 2017; Khoshsima & Hashemi Toroujeni, 2017b; OECD, 2010; Wilson, Genco, & Yager, 1985). For example, OECD (2010) reported that there were no difference in test performance between computerized and PPT versions of tests among participants (n = 5,878) from Denmark, Iceland and Korea. In a review of educational and psychological measurement approaches, Bunderson, Inouye

& Olsen (1989) reported that 48% of previous studies showed no difference between the two testing modes in test performance, 13% of studies showed the superiority of computerized test and 39% of studies showed that PPT was superior. The concept of equivalence was supported by only nearly half of the studies, and the differences were ascertained in achievement tests such as science, language and mathematics tests, and also obviously in psychological tests such as personality and neuropsychological assessment (e.g. Friedrich & Bjornsson, 2008; Choi, Kim & Boo, 2003; DeAngelis, 2000).

To assess the testing mode administration effects on scores obtained from two versions of the same test and consequently on the performance of test takers become inevitable when more conventional tests are converted to computerized administration. Since the reduction of test mode effects is necessary and beneficial to test practitioners due to the desirability of accurate measurement rather than inaccurate one, extensive body of research with mixed results conducted on the comparability of test modes have being done.

Converting a traditional paper and pencil test to the computerized test resulted into two types of linear and adaptive testing strategies. In CALT, not only the medium of administration transforms from paper to computer, but also the test algorithm turns from linear to adaptive which allows the possibility of presentation and administration of test items tailored to each test taker's ability. Therefore, in comparability studies of traditional paper and pencil testing and computer adaptive language testing, not only the administration mode, but also paradigm effect on test takers' performance can be studied to ensure the comparability of CALT and its PPT counterpart. The administration mode effect has been widely examined in comparability study of PPTs and CBTs.

In some cases, the mode and paradigm effects are dumbfounded with each other in comparability study of CALT and its PPT counterpart. Then, to solve this problem, some studies separate testing mode effects and paradigm effects by comparing linear CBT and CALT to examine just the paradigm effect of testing on test takers performance. Such a comparability study between linear CBT and CALT on the three GRE measures was done by Schaeffer, Steffen, Smith, Mills, and Durso (1995). In this study, two examinations were done. The first one examined the comparability of scores obtained from two different testing paradigms including linear computer based and computer adaptive versions of the three GRE General Test measures. It was found that comparable scores to CBT counterpart were produced by verbal and quantitative CALTs. But the scores produced by analytical CALT were found not to be comparable to the analytical CBT scores.

However, an additional examination was done to show the differences in analytical CALT and CBT scores due to the testing paradigm difference. It was found that the large differences between analytical CALT and CBT scores required an adjustment. Therefore, in order to enhance the comparability of analytical CALT and CBT scores, the analytical CALT was equated to the analytical CBT. This equating provided new analytical CALT conversions that resulted in comparable analytical CALT and CBT scores. They found that analytic CALT and CBT produced incomparable scores which were in favour of the CALT while both versions were comparable for the other two measures. A comparability study of CBT and CALT forms of a vocabulary test to measure the efficiency of tests was done by Vispoel, Rocklin, and Wang (1994). The findings of this study show that CALT version of the test produced more exact and accurate ability estimates and fewer items were administered in CALT in order to reach the same accuracy as the fixed-item CBT. Similar results were also found by Vispoel, Wang, and Bleiler (1997) when they compared two versions of tests in assessing music listening skills.

The results of a meta-analysis of a research conducted by Wang et al., in 2008 showed that the performance of K-12 students on mathematic achievement test was not considerably influenced by testing mode of administration. But it was found that the paradigm of testing including linear or adaptive algorithm (applied to CBT and CALT, respectively) to deliver the test introduced in the study as the moderator variable had significant effect on test takers' performance on math test (Wang et al., 2007, 2008). Linear test scores discrepancies were larger than the scores obtained from the adaptive test in the study. Although the analyses of students' obtained scores from mathematics conventional paper and pencil test and its computerized counterpart showed that some variables such as study design, grade level, sample size, type of test, computer delivery method, and computer practice caused no significant discrepancies in test performance, they reached the conclusion that test paradigm is one of the significant influencing factors that lead to considerable incomparability between the CALT and PPT versions of a test.

In another comparability study of paper and pencil and computer adaptive test scores on the GRE general test done by Schaeffer et al., in 1998 it was found that mean scores of each measure (verbal, quantitative, and analytical) on CALT were higher than mean scores on PPT. The investigation proved that CALT test takers who did not complete their CALTs obtained higher mean scores than would be predicted. However, mean scores for test takers who completed their CALTs were similar to mean scores for PPT test takers. The results were obtained using a new psychometrically-defensible

CALT scoring method called proportional scoring. It is also hypothesized that the increased average time per item may be one of the factors that lead to the higher scores. However, it seems that CALT and PPT scores are comparable when CALT test takers answer all of the items in the test (Schaeffer, Bridgeman, Smith, Lewis, Ptenza, & Steffen, 1998).

Another comparability study of computer adaptive testing and its conventional counterpart known as ASVAB (Armed Services Vocational Aptitude Battery) was done by Moreno *et al.* in 1983. In fact the Navy Personnel Research and Development Centre of U.S Army that was under the supervision of Department of Defence was trying to replace its paper-and-pencil Armed Services Vocational Aptitude Battery including a fixed sequence of test questions given to all test takers known as ASVAB with a computerized adaptive test (CAT) due to its capability to tailor the items of aptitude test to every test taker by selecting those items whose psychometric characteristics match his/her apparent ability and knowledge level. The study sought to determine the reliability of two test versions and the relationship between selected paper-and-pencil ASVAB subtests and their CAT counterparts which contained three subtests constructed to measure Arithmetic Reasoning (CATAR), Word Knowledge (CATWK), and Paragraph Comprehension (CATPC). Fixed length design was selected for the CAT subtests which Bayesian sequential tailored testing procedure (Owen, 1969, 1975) have been used for to optimize a mathematical function of the difference between the examinee's estimated ability and the item's difficulty. Three tests including initial ASVAB test, ASVAB retest and CAT test were administered to subjects and the results showed that CAT subtest scores correlated as highly with ASVAB initial test scores as did the ASVAB retest scores even though the CAT subtests included only half number of conventional test items and ability estimates from CAT subtests loaded on the same factors as did their counterpart ASVAB subtests (Moreno, Wetzel, McBride & Weiss, 1983).

Similar studies investigated the effect of some so called moderator factors such as prior computer experience and familiarity on test takers performance (Hashemi Toroujeni, 2016). The researcher found positive attitudes toward computers related to more use of computers. It was also found that computer familiarity had no significant effect on test takers performance and their willingness to take the computerized test when two versions of the same test were available. The Guidelines for Computer-Based Tests and Interpretations (APA, 1986) recommended eliminating the possible effects of some moderator variables such as computer experience on test scores and testing takers performance. Of course, some other studies examined the effects of test anxiety caused

by unfamiliarity with the testing environment, and other factors such as inflexible software and computer anxiety. They found that these moderator factors affect test results negatively and adversely (Kveton, Jelinek, Voboril, & Klimusova, 2007; Smith & Caputi, 2007). Since some students bring up unfamiliarity with computerized mode of testing as the main reason of their falling in this kind of testing and complain that their computerized test score is not the real representative of their language proficiency, the necessity of studying the prior frequent use of computer as a moderator variable in CALT have to be considered.

About the relationship of computer familiarity as the frequently cited contributor to score differences with the examinee performance on both forms of testing, Wallace and Clariana (2000) said that learner characteristics such as computer experience were associated with higher post-test performance for computerized test (in their case, web-based test). They found out that lower ability learners were less familiar with computers. In another study, although findings revealed the priority of CBT over PPT with .01 degree of difference at $p<05$, it was indicated that two prior computer familiarity and attitudes external moderator factors had no significant effect on test takers' CBT scores (Khoshsima & Hashemi Toroujeni, 2017a).

Watson (2001) also reported that although there was no relationship between age and sex with students' performance, students with higher academic attainment and those with greater frequency of computer use benefited mostly from computer based instruction. In addition, some other studies showed that students with a good knowledge of computer use fell more free and comfortable to utilize computerized kind of testing (O'Malley, Kirkpatrick, Sherwood, Burdick, Hsieh, & Sanford, 2005; Poggio, et al., 2005). Prior computer experience variable can be introduced as one of the most critical reason causing discrepancies in test mode performance (Taylor, Kirsch, Eignor, and Jamieson, 1999). Some indefinite conclusions concerning to the impact of computer familiarity on performance were resulted from other studies. One of the major reasons of converting the paper based TOEFL test into IBT version was the incapability of paper based version of multiple choice test to measure the higher order processing skills that are usually employed in constructing and communicating meaning (Lynch, 2000). However, the Educational Testing Service (ETS) stated that the greatest danger to the proposed improved validity of the CBT version was the effect of computer familiarity (Kirsch, Jamieson, Taylor, & Eignor, 1998). The resulted findings of an examination done by Kirsch et al. that studied the relationship between levels of computer familiarity and performance on the computerized TOEFL test after implementation of an online familiarization training showed no significant difference between prior

computer use and experience of test takers and their performance on the computerized test (Kirsch et al., 1998). However it is likely, as with the Powers and O'Neill (1992) study, that either pretest computer training negated the low pre-familiarity levels of examinees, or computer familiarity may have played only a small part in performance, as it does not appear to have the significant impact once assumed. Regarding the former point, a number of authors have suggested that if computer familiarity is a key factor associated with the test mode effect, it may be rapidly diminishing with increased access to computers in schools and the home (Clariana & Wallace, 2002; Kirsch et al., 1998; Lynch, 2000; McDonald, 2002). Besides computer familiarity and computer attitude, testing mode and paradigm preferences of test takers that are typically related to high stakes standardized test administration are being noticed in recent researches. For example, some studies reached the conclusion that test takers preferred the computer form of the test (Pinsoneault, 1996; Hansen, *et al*., 1997; Vispoel, 2000; Vispoel *et al*., 2001). Some studies have also shown that computer anxiety, lack of confidence, and lack of enjoyment influence both the acceptance of computers and their use as a teaching and learning tool (Gressard & Loyd, 1986; Smith & Kotrlik, 1990; Woodrow, 1991; Fletcher & Deeds, 1994). McDonald (2002) reported that computer aversion or anxiety refers to the unpleasant feeling of fear and uneasiness experienced by student when s/he is interacting with a computer or anticipating an interaction (p.305). Some studies suggest that computer aversion overlaps with computer experience construct and the hypothesis that computer aversion results from lack of computer familiarity is reinforced (Levine & Donitsa-Schmidt, 1998). Although Durndell and lightbody (1994) found out that there is not any inverse relationship between computer familiarity and computer aversion, in another investigation of the published studies of computer aversion from 1990 to 1996 done by Chua *et al.*, a conflicting conclusion has been reached and it was reported that computer aversion was inversely related to computer familiarity and use (Chua, Chen, and Wong, 1999).

2.1 Developing Computer Adaptive Testing (CAT)

Compute-adaptive test (CAT) is a subtype and subtest of computer-assisted language test because it is administered at computer terminal or on personal computer. The computer-adaptive subtype of computer-assisted or computer-based tests has three additional characteristics: (a) the test items are selected and fitted to the individual students involved, in other words test items are tailored based on the individual test taker's ability and level of knowledge (b) the test is terminated when the knowledge and ability level of individual test taker is specified, and, as a consequence, (c) computer

adaptive tests are usually relatively short in terms of the number of items involved and the time needed, so CAT leads to save time of test administration (Madson, 1991; Wainer, 1990).

Unlike the conventional fixed-length paper-pencil tests, this flexi-level strategy provide the situations and conditions for test-takers in which they answer just the questions appropriate to their proficiency levels and the need to answer numerous difficult or easy questions is eliminated. In fact, as Madson (1991) puts it, "*the computer-adaptive test (CAT) is uniquely tailored to each individual*" (p.237). Then, one unique feature of computer-adaptive test is that test taker takes the test that is appropriate and suitable to his/her own particular ability level and the test is automatically terminated when the examinee's ability level has been located.

Tung (1986) elucidated well how to develop a computer-adaptive test (CAT). Development and implementation of computer-adaptive mode of testing is in its initial stages. The well-known implications and advantages of CAT including efficiency, flexibility in administration time and item selection, security issues, quicker availability of the results and scoring accuracy have been leading it into more popularity among test practitioners.

The issue that currently needs more attention and prompt investigation of researchers is to study the testing mode and paradigm effects on comparability and equivalency of the data obtained from two modes of presentation, i.e. traditional paper-and-pencil (PPT) and computerized tests. According to Chalhoub-Devil and Devil (1999), comparability researches and studies in second language tests are in short supply, and he also emphasized over the importance of conducting comparability studies in local settings to detect any potential test-delivery-medium effect when a traditional PPT test is converted to a computerized one. To establish comparability and equivalency of computerized test with its paper-and-pencil counterpart is of importance and critical. Research has focused on the equivalency of computer and paper-administered tests in terms of scores (Choi, Kim, and Boo, 2003; Kenyon and Malabonga, 2001). Recently, some studies have been done to indicate that in order to replace computer-adaptive test with conventional paper-and-pencil one, we need to prove that these two versions of test are comparable, in other words the validity and reliability of computerized counterpart are not violated, but there is no agreed upon theoretical explanation for the test mode effects. The comparability is achieved through equivalent scores of two test versions.

3. Key Factors related to CAT

Individual characteristics of test takers may provide a cornerstone and groundwork for a theory explaining the foundational aspects involved in test performance in two different testing modes with different paradigms. Inevitable questions about test takers' reactions to and attitudes about computerized version of paper-and-pencil test are raised after the introduction of the worldwide computerized version of the Test of English as a Foreign Language to evaluate general English proficiency of those whose native language is not English. Due to the probable impact of these issues on test taking motivation, test performance and thereby on test validity, these issues are of prime importance (Ryan & Ployhart, 2000). About the influence of prior computer familiarity on test performance, Taylor, Kirsch, Eignor, and Jamieson (1999) claimed that computer experience is related to performance on the paper-based version of the TOEFL. They demonstrated that those who obtained high scorers were more familiar with computers. About the test takers' gender, according to them, men were more familiar, and also, Spanish speakers were more familiar than Japanese speakers who were less familiar.

Some other researches conducted in academic settings with adult participants demonstrated that computer familiarity is related to acceptance and other attitudes about computers (Powers & O'Neill, 1993; Wilder, Mackie, & Cooper, 1985), anxiety about computers (Kernan & Howard, 1990; Powers & O'Neill, 1993), and attitudes about computerized tests (Burke, Normand, & Raju, 1987). Of course less is known about the relationship of familiarity and computer anxiety with performance on computer-based tests. Familiarity was related to performance in one study (Lee, 1986) but not in three others (Powers & O'Neill, 1993; Taylor et al., 1999; Wise, Barnes, Harvey, & Plake, 1989), and anxiety and performance were unrelated in three studies (Mazzeo, Druesne, Raffeld, Checketts, & Muhlstein, 1992; Powers & O'Neill, 1993; Wise et al., 1989). Finally, several studies have established high acceptance of computerized tests (Powers & O'Neill, 1993; Schmidt, Urry, & Gugel, 1978; Schmitt, Gilliland, Landis, & Devine, 1993). And Jamieson, Taylor, Kirsch, and Eignor (1999) found that providing TOEFL test takers with a computer-administered tutorial on taking a prototype computer-based TOEFL increased their acceptance of that test, particularly among those who were less familiar with computers. Although, attitudes about admission test seemed to be generally negative in two countries, attitudes about computer-based TOEFL appeared to be relatively positive in the countries in which the study has been done (Stricker & Wilder & Rock, 2004).

Some factors that determine the attitudes towards the use of computer in testing setting are based on computer familiarity, knowledge level, skills and abilities, ease of access to computer, formal computer training, gender and some else. In a study done by Wallace and Clariana (2000), learners' characteristics including learner ability, computer familiarity, and non-competitiveness related to the higher performance of posttest for the group which web-based exam was administered for were investigated. The central findings suggest that learners in lower ability were less familiar with computers, and competitive learners did not do as well online rather they did well in traditional classroom setting. Similarly, Watson (2001) reported that students with higher academic attainment and also those with greater frequency of computer use benefited most from computer-based instruction, while age and gender were not factors. Gender and age were studied in other research (Parshall & Kromery, 1993). Then, more to the point, learner characteristics have been directly associated with test mode effect and the present research studies computer attitude, prior testing mode preference and computer familiarity, computer anxiety, gender and age factors based on two testing modes and paradigms performance.

Attractive test delivery via computers is being prevalent due to the existence of current relatively cheap but powerful microcomputers. Despite clear evidence of the digital divide that exists in poor, urban, minority schools, many schools and districts insist on administrating high stakes tests via computers (Thomas, 2008, pp. 4-6). In spite of vigorous the insistence on administering high or low stakes tests via computers in many contexts, some problems yet exist in academic contexts related to the use of technology to assess learning process of students and a whole range of issues have yet to be resolved. Since test-takers performance and consequently their future life may be influenced by the consequent effects of two testing administration modes and paradigm, it is necessary to assess whether tests with two versions and paradigms are reliable and valid, whether they are comparable and equivalent. Some studies have been conducted with adult examinees to evaluate the comparability of scores obtained from computerized and paper-pencil versions of a test to measure the effect of administration mode (Mead & Drasgow, 1993; Hetter, Segall & Bloxom, 1997). To make sure of the consistency and fairness of the test results based on validity and reliability factors is the aim of the comparability study. Reliability, according to Bachman and Palmer (1996), is a crucial aspect in test usefulness and is worth to do research on. In fact, comparability of CALT and PPT can be evaluated according to some general categories of criteria including (1) validity (construct and predictive), (2) psychometric (such as reliability that can be examined at both test and item levels), and (3) statistical

assumption/test administration mentioned by Wang and Kolen (2001). Of course, due to some critical challenging issues relating to administration mode of CALT such as item parameter estimation, item selection method, item scoring procedures, and the stopping rule that make it different from PPT (Paper-and-Pencil Based) or CBT (Computer-Based Testing), the evaluation procedures may become more complicated (Green, Bock, Humhpreys, Linn, & Reckase 1984). In fact, before introducing a computerized version of a test, it is necessary to determine whether scores from the computerized and paper-and-pencil versions can be used interchangeably. Unless comparability is established between computerized and paper-and-pencil versions of a test, the scores obtained through a computerized version cannot be interpreted in the same way as scores from a paper-and-pencil version. As noted in the Guidelines for Computer-Based Tests and Interpretations (APA, 1986), "*When interpreting scores from the computerized versions of conventional tests, the equivalence of scores from computerized versions should be established and documented before using norms or cutting scores obtained from conventional tests*" (P. 18) . Applying norms and standards without empirically established comparability between two versions of a test could result in unfair and inappropriate decisions about individuals.

The Guidelines for Computer-Based Tests and Interpretations (APA, 1986) emphasizes that the comparability of scores from computerized and paper-and-pencil versions of a test should be studied empirically. To evaluate the comparability of a computerized test to a corresponding paper-and-pencil test, psychometric properties need to be examined first. Two tests can be considered psychometrically comparable if they produce scores with similar rank orderings, distributions, and correlations with other variables. Also, computer-related factors such as computer anxiety, computer experience, computer attitude, gender and age need to be examined to determine the extent to which they affect the comparability of scores between the computerized and paper-and-pencil tests.

Another issue that needs to be clarified in a PBT and CBT comparability study, as raised by Wise and DeMars (2003) is motivational factors which might also have an impact on test performance. Wise and DeMars pointed out that regardless of how much psychometric care is applied to test development, or how equal the testing modes are, to the extent that test takers are not motivated to respond to the test (e.g. due to low efficacy or boredom), test score validity will be compromised. The test taker motivation model (Pintrich, 1989) specifies that the effort test takers will direct towards a test is a function of how well they feel they will do on the test, how they perceive the test to be, and it related to their affective reactions regarding the test. This is the theoretical model

that underlies the relationship among motivation, testing mode and test performance. Besides that, the self-determination theory (Wenemark, Persson, Brage, Svensson & Kristenson, 2011) states that increases test-takers' motivation will increase the willingness to take the test or response rates, and thus it will enhance learning. Therefore, testing motivation is an aspect worth investigating in testing mode comparability studies because it can pose a threat to the validity of inferences made regarding assessment test results (Shuttleworth, 2009).

As a result, the current researcher decided to investigate the comparability of Computer-Adaptive Language Test and Paper-And-Pencil Test based on reliability and validity aspects of testing and mode and paradigm effects of two various versions of tests on test takers' performance in alleviating EFL learners' assessment dilemmas. Therefore, in this comparability study, both the administration mode and paradigm effects on examinees' performance were studied to ensure the comparability of the CALT and its PPT counterpart.

Larson and Madsen (1985) defined CALT program as a stimulus for test developers and test practitioners to develop and design various kinds of computer adapted tests 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 throughout the 1990s (e.g., Kaya-Carton, Carton & Dandonoli, 1991; Burston & Monville-Burston, 1995; Brown & Iwashita, 1996; Young, Shermis, Brutten & Perkins, 1996). The standards for developing computerized testing to administer and replace with its paper-and-pencil counterpart requires that equivalent test scores be established for the paper-and-pencil based testing (PPT) and computerized adaptive testing (CAT). Although the two testing modes are nearly identical in most comparability studies, significant discrepancies of test scores are observed. Therefore, the validity of replacing CAT with PPT in educational assessment in academic contexts is under question. Then, as the first step to replace a CAT program with PPT test, mode and paradigm effects of two versions of tests on test takers' performance should be investigated to see whether the two sets of scores are comparable and consequently valid or reliable. And it is important to see whether the scores derived from a CAT measure had similar characteristics to scores derived from a linear fixed-length PPT.

4. The Necessity for Comparability Study

Testing in education is a key component of learning experience that attempts to measure learners' knowledge, intelligence, or other characteristics in a systematic way. As computerized testing has become extremely prolific in the last 10-15 years (Hashemi Toroujeni, 2016), the amount of industry research on the comparability between paper-based and computer-based exams has grown considerably. In fact the advantages of using computers in language testing have been leading many organizations, institutes, universities and others to move eagerly toward computerized version of tests. There are two major kinds of computerized testing strategies: 1) fixed length linear conventional tests that are constructed by selecting a fixed set of items for administration to a group of individuals. 2) Adaptive tests that are efficient even for a group of individuals who are widely different in ability. This type of testing paradigm is used based on a simple concept: more information can be obtained from a test item if the item is matched to the ability level of the examinee.

Since evaluating the comparability of paper-based and computer-based tests is crucial before introducing computer aided assessment into any context, the purpose of the current study was to compare students' performance between PPT and CALT versions of the tests. The study focused on the comparability and equivalency of the product of the tests i.e. scores and the processes used to achieve that product. Chalhoub-Deville and Devil (1999) pointed out that there is a scarcity of comparability research on localized language tests needed to detect any potential impact of the test delivery mode when converting conventional paper tests to computerized tests.

Scores from a test should reflect differences among individuals only in characteristics relevant to what the test is supposed to measure. Hofer and Green (1985), however, counted test-taker's computer anxiety and computer familiarity among the reasons for incomparability of scores between computerized and paper-and-pencil versions of a test. Mazzeo and Harvey (1988) also reported that computer familiarity might affect scores. Kolen (1996) wrote that "*scores on a paper-and-pencil and a computerized test might be comparable for examinee groups with considerable computer experience, but not for examinee groups with little computer experience*" (P. 7).

5. Conclusion

Converting a traditional paper and pencil test to the computerized test resulted into two types of linear and adaptive testing strategies. In CALT, not only the medium of

administration transforms from paper to computer, but also the test algorithm turns from linear to adaptive which allows the possibility of presentation and administration of test items tailored to each test taker's ability. Therefore, in comparability studies of traditional paper and pencil testing and computer adaptive language testing, not only the administration mode, but also paradigm effect on test takers' performance can be studied to ensure the comparability of CALT and its PPT counterpart. The administration mode effect has been widely examined in comparability study of PPTs and CBTs.

In some cases, the mode and paradigm effects are dumbfounded with each other in comparability study of CALT and its PPT counterpart. Then, to solve this problem, some studies separate testing mode effects and paradigm effects by comparing linear CBT and CALT to examine just the paradigm effect of testing on test takers performance. Such a comparability study between linear CBT and CALT on three GRE measures was done by Schaeffer, Steffen, Smith, Mills, and Durso (1995). In this study, two examinations were done. The first one examined the comparability of scores obtained from two different testing paradigms including linear computer based and computer adaptive versions of the three GRE General Test measures. It was found that comparable scores to CBT counterpart were produced by verbal and quantitative CALTs. But the scores produced by analytical CALT were found not to be comparable to the analytical CBT scores. However, an additional examination was done to show the differences in analytical CALT and CBT scores due to the testing paradigm difference. It was found that the large differences between analytical CALT and CBT scores required an adjustment. Therefore, in order to enhance the comparability of analytical CALT and CBT scores, the analytical CALT was equated to the analytical CBT. This equating provided new analytical CALT conversions that resulted in comparable analytical CALT and CBT scores. They found that analytic CALT and CBT produced incomparable scores which were in favour of the CALT while both versions were comparable for the other two measures (Schaeffer, Steffen, Smith, Mills, and Durso, 1995). Then, in every comparability study, it should be examined that whether students' performance on achievement test differs by mode (text versus digital) and paradigm (linear versus adaptive).

Furthermore, factors such as comparability, equivalency, reliability and validity of two versions of test should be examined before introducing CALT. Additionally, the other factors that are worth considering in comparability studies include computer anxiety, prior computer attitude, prior testing mode and paradigm preference and computer familiarity and experience, gender and age factors as the major highly

influencing test takers' characteristics on their performance. Then post-test data and learner-self-report information obtained from empirical researches can help identify some key factors that relate to the test mode and test paradigm effects.

An illustration of the subject of current research and comparability study between CAT, CBT and PPT as a necessity in language testing domain has been presented in this paper. This paper also emphasized on examining the relationship between some external variables and test takers' performance. Several related studies have been mentioned in this paper too. Some approaches of investigating score equivalency in CAT, CBT and PPT and the association between some test mode factors have been discussed. The researcher reviewed some related studies and the findings of some researchers have been shown. For example, some studies found significant difference between the scores obtained from two CBT and PPT versions (Pomplun et al., 2002; Choi et al., 2003), and in some other studies no significant difference was found (Russell & Haney, 1996; Pommerich, 2004). To compare test scores received from CAT, CBT and PPT and to examine the relationship between some external factors such as computer familiarity, computer attitudes, computer aversion or anxiety and preference of testing mode that may influence the performance of test takers are the main goals of comparability studies. Those important factors that were addressed in the current study have been studied in many researches and it is still recommended that the future comparability studies investigate them more especially in local settings to apply the findings and results practically. It is also suggested that the comparability studies consider gender difference and investigate performance difference between male and female participants. The researcher hopes that the present study adds to the available knowledge in comparability study of CAT and PPT.

References

1. Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole.
2. American Psychological Association (APA). (1986). *Guidelines for computer-based tests and interpretations.* Washington, DC: Author.
3. Bachman, L.F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

4. Bachman, L.F., Davidson, F., Ryan, K. and Choi, I.-C. (1995): *Studies in language testing 1: an investigation into the comparability of two tests of English as a Foreign Language*. Cambridge: Cambridge University Press.
5. Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.
6. Bayroff, A. G. (1964). *Feasibility of a programmed testing machine* (U.S. Army Personnel Research Office Research Study 6403). Washington, DC: U.S. Army Behavioural Science Research Laboratory.
7. Bayroff, A. G., Thomas, J. J., & Anderson, A. A. (1960). *Construction of an experimental sequential item test* (Research Memorandum 60-1). Washington, DC: Department of the Army, Personnel Research Branch.
8. Becker, Kirk A. & Bergstrom, Betty A. (2013). Test administration models. *Practical Assessment, Research & Evaluation*, 18(14). Available online: http://pareonline.net/getvn.asp?v=18&n=14.
9. Bejar, I. I., & Braun, H. (1994). *On the synergy between assessment and instruction: Early lessons from computer-based simulations*. Machine-Mediated Learning, 4, 5-25.
10. Bennett, R. E. (1999). *How the Internet will help large-scale assessment reinvents itself*. Education Policy Analysis Archives, 9(5), 1-25.
11. Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). *Altering the level of difficulty in computer adaptive testing*. Applied Measurement in Education, 5(2), 137-149. https://doi.org/10.1207/s15324818ame0502_4.
12. Binet, A., & Simon, Th. A. (1905). *Méthode nouvelle pour le diagnostic du niveau intellectuel des anormaux*. L'Année Psychologies, 11, 191–244. https://doi.org/10.3406/psy.1904.3675.
13. Brown, A. & Iwashita, N. (1996). *The role of language background in the validation of a computer- adaptive test*. System, 24(2), 199-206. https://doi.org/10.1016/0346-251X(96)00004-8.
14. Bugbee Jr., A. C. (1996). *The equivalence of paper-and-pencil and computer-based testing.* Journal of Research on Computing in Education, 28, 282–299. https://doi.org/10.1080/08886504.1996.10782166.
15. Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). *The four generations of computerized educational measurement*. In R. L. Linn (Ed.), Educational Measurement (pp. 367–407). Washington, DC: American Council on Education.
16. Burke, M. J., Normand, J., & Raju, N. S. (1987). *Examinee attitudes toward computer administered ability tests.* Computers in Human Behaviour, 3, 95–107. https://doi.org/10.1016/0747-5632(87)90015-X.

17. 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.
18. Canale, M. (1983). *On some dimensions of language proficiency*. In Oller, J.W. Jr., editor, Issues in language testing research. Rowley, MA: Newbury House, 333–42.
19. Chalhoub-Deville, M., & Deville, C. (1999). *Computer adaptive testing in second language contexts.* Annual Review of Applied Linguistics, *19,* 273-299. https://doi.org/10.1017/s0267190599190147.
20. Challoner, J. (2009). *1001 Inventions that changed the world* (Cassell Illustrated: 2009). 3
21. Chapelle, C. A. (2010). *Technology in language testing* [video]. Retrieved November 14, 2012 from http://languagetesting.info/video/main.html.
22. Choi, I.-C. (1991). *Theoretical studies in second language acquisition: application of item response theory to language testing*. New York: Peter Lang Publishing.
23. Choi, I.-C., Kim, K.S., and Boo, J. (2003). *'Comparability of a paper-based language test and a computer-based language test'*, Language Testing 20(3), 295–320. https://doi.org/10.1191/0265532203lt258oa.
24. Chua, S. L., Chen, D. T., & Wong, A. F. L. (1999). *Computer anxiety and its correlates: A meta-analysis.* Computers in Human Behavior, 15, 609–623. https://doi.org/10.1016/S0747-5632(99)00039-4.
25. Clariana, R., & Wallace, P. (2002). *Paper-based versus computer-based assessment: Key factors associated with the test mode effect*. British Journal of Educational Technology, 33, 593-602. https://doi.org/10.1111/1467-8535.00294.
26. Cummins, J.P. (1983). *Language proficiency and academic achievement*. In Oller, J.W. Jr., editor, Issues in language testing research. Rowley, MA: Newbury House, 108–26. 110.
27. DeAngelis, S. (2000). *Equivalency of computer-based and paper-and-pencil testing.* Journal of Allied Health, 29(3), 161–164.
28. Durndell, A., & Lightbody, P. (1994). *Gender and computing: Change over time?* Computers & Education, 21, 331–336. https://doi.org/10.1016/0360-1315(93)90036-I.
29. Fletcher, W. E. & Deeds, J. P. (1994). *Computer anxiety and other factors preventing computer use among United States secondary agricultural educators*. Journal of Agricultural Education, 35(2), 16-21. https://doi.org/10.5032/jae.1994.02016.

30. Friedrich, S., & Bjornsson, J. (2008). *The transition to computer-based testing – New approaches to skills assessment and implications for large-scale testing*. http://crell.jrc.it/RP/reporttransition.pdf (accessed May 23, 2011).
31. Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). *Technical guidelines for assessing computerized adaptive tests.* Journal of Educational Measurement, 21(4), 347-360. https://doi.org/10.1111/j.1745-3984.1984.tb01039.x.
32. Gressard, C. P., & Loyd, B. H. (1986). *Validation studies of a new computer attitude scale.* Association for Educational Data Systems Journal, 18(4), 295-301. https://doi.org/10.1080/00011037.1986.11008443.
33. Hansen, J.-I.C., et al., (1997). *Comparison of user reaction to two methods of Strong Interest Inventory administration and report feedback.* Measure and Evaluation in Counseling and Development, 30, 115–127.
34. Hashemi Toroujeni, S.M. (2016). *Computer-Based Language Testing versus Paper-and-Pencil Testing: Comparing Mode Effects of Two Versions of General English Vocabulary Test on Chabahar Maritime University ESP Students' Performance*. Unpublished thesis submitted for the degree of Master of Arts in TEFL. Chabahar Marine and Maritime University (Iran) (2016).
35. Hetter, R. D., Segall, D. O., & Bloxom, B. M. (1997). Evaluating item calibration medium in computerized adaptive testing. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 161–167). Washington, DC: American Psychological Association. https://doi.org/10.1037/10244-017.
36. Hofer, P., & Green, B. (1985). The challenge of competence and creativity in computerized psychological testing. *Journal of Consulting and Clinical Psychology,* 53, 826- 838. https://doi.org/10.1037/0022-006X.53.6.826.
37. International Test Commission. (2004). *International Guidelines on Computer-Based and Internet- Delivered Testing*. Retrieved January 21, 2011 from http://www.intestcom.org/itc_projects.htm.
38. Jamieson, J., Taylor, C., Kirsch, I., & Eignor, D. (1999). *Design and evaluation of a computer-based TOEFL tutorial* (TOEFL Research Report 62; ETS Research Report 99-01). Princeton, NJ: Educational Testing Service.
39. 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.

40. Kenyon, D.M. and Malabonga, V. (2001). 'Comparing *examinee attitudes toward computer-assisted and other oral proficiency assessments'*, Language Learning and Technology 5(2), 60–83.
41. Kernan, M. C., & Howard, G. S. (1990). *Computer anxiety and computer attitudes: an investigation of construct and predictive validity issues.* Educational and Psychological Measurement, 50, 681–690. https://doi.org/10.1177/0013164490503026.
42. Khoshsima, H. & Hashemi Toroujeni, S.M. (2017a). *Transitioning to an Alternative Assessment: Computer-Based Testing and Key Factors related to Testing Mode.* European Journal of English Language Teaching, Vol.2, Issue.1, pp. 54-74, February (2017). ISSN 2501-7136. http://dx.doi.org/10.5281/zenodo.268576.
43. Khoshsima, H. & Hashemi Toroujeni, S.M. (2017b). *Comparability of Computer-Based Testing and Paper-Based Testing: Testing Mode Effect, Testing Mode Order, Computer Attitudes and Testing Mode Preference.* International Journal of Computer (IJC), (2017) Volume 24, No 1, pp 80-99. ISSN 2307-4523 (Print & Online), http://ijcjournal.org/index.php/InternationalJournalOfComputer/article/view/825/4188.
44. Khoshsima, H., Hosseini, M. & Hashemi Toroujeni, S.M. (2017). *Cross-Mode Comparability of Computer-Based Testing (CBT) versus Paper and Pencil-Based Testing (PPT): An Investigation of Testing Administration Mode among Iranian Intermediate EFL learners.* English Language Teaching, Vol.10, No.2; January (2017). ISSN 1916-4742 (Print), ISSN (1916-4750). http://dx.doi.org/10.5539/elt.v10n2p23.
45. Kingsbury, G. G., & Zara, A. (1989). *Procedures for selecting items for computerized adaptive tests.* Applied Measurement in Education, 2, 359-375. https://doi.org/10.1207/s15324818ame0204_6.
46. Kirsch, I., Jamieson, J., Taylor, C., & Eignor, D. (1998). *Computer familiarity among TOEFL examinees (TOEFL Research Report 59).* Princeton, NJ: Educational Testing Service.
47. Kolen, M. J. (1996). *Threats to score comparability with applications to performance assessments and computerized adaptive tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.
48. Kranzberg, M and Davenport, W.H (1972). *Technology and Culture: An Anthology* (New York: Meridian).
49. Kveton, P., Jelinek, M., Voboril, D., & Klimusova, H. (2007). *Computer-based tests: the impact of test design and problem of equivalency.* Computers in Human Behavior, 23(1), 32-51. https://doi.org/10.1016/j.chb.2004.03.034.

50. Larson, J. W. & Madsen. H. S. (1985). *Computer-adaptive language testing: Moving beyond computer-assisted testing.* CALICO Journal, 2(3), 32-6.
51. Larson, J. W. & Madsen. H. S. (1989). "*S-CAPE: A Spanish Computerized Adaptive Placement Exam.*" Modern Technology in Foreign Language Education: Application and Projects, edited by F. Smith. Lincolnwood, IL: National Textbook.
52. Lee, J. A. (1986). *The effects of past computer experience on computerized aptitude test performance.* Educational and Psychological Measurement, 46, 727–733. https://doi.org/10.1177/0013164486463030.
53. Levine, T., & Donitsa-Schmidt, S. (1998). *Computer use, confidence, attitudes, and knowledge: A causal analysis.* Computers in Human Behavior, 14, 125–146. https://doi.org/10.1016/S0747-5632(97)00036-8.
54. Lord, F. M. (1970). *Some test theory for tailored testing*. In W. H. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance (pp. 139–183). New York: Harper & Row.
55. Lord, F. M. (1971a). *Tailored testing, an approximation of stochastic approximation*. Journal of the American Statistical Association, 66, 707–711. https://doi.org/10.1080/01621459.1971.10482333.
56. Lynch, R. (2000). *Computer-based testing: The test of English as a foreign language (TOEFL).* The Source, Fall 2000. Retrieved January 6, 2004, from http://www.usc.edu/dept/education/ The Source/>Fall2000
57. MacDonald, A. S. (2002). *The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments.* Computers & Education, 39, 299-312. https://doi.org/10.1016/S0360-1315(02)00032-5.
58. Madsen, H. S. (1991). *Computer-adaptive test of listening and reading comprehension: The Brigham Young University approach*. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 237-257). New York: Newbury House.
59. Manip Ther, J.M., (2010). *Mode of Administration Bias*. The Journal of Manual.
60. Mason, B. J., Patry, M., & Berstein, D. J. (2001). *An examination of the equivalence between non-adaptive computer based and traditional testing.* Journal of Educational Computing Research, 24(l), 29-39. https://doi.org/10.2190/9EPM-B14R-XQWT-WVNL.
61. Mazzeo, J., Druesne, B., Raffeld, P., Checketts, K., & Muhlstein, A. (1992). *Comparability of computer and paper-and-pencil scores for two CLEP general*

examinations (College Board Report 91-5; ETS Research Report 92-14). New York: College Entrance Examination Board.

62. Mazzeo, J., & Harvey, A.L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests* (College Board Report No. 88-8). New York: College Entrance Examination Board.
63. Mead, A. D., & Drasgow, F. (1993). *Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis*. Psychological Bulletin, 114, 449-458. https://doi.org/10.1037/0033-2909.114.3.449.
64. Moreno, K. E., Wetzel, C. D., McBride, J. R., & Weiss, D. J. (1983). *Relationship between corresponding Armed Services Vocational Aptitude Battery and computerized adaptive testing subtests*. NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER, San Diego, California 92152
65. Muckle, T. J., Bergstrom, B. A., Becker, K., & Stahl, J. A. (2008). *Impact of altering randomization intervals on precision of measurement and item exposure.* Journal of Applied Measurement, 9(2), 160-167.
66. Noijons, J. (1994). Testing computer assisted language tests: Towards a checklist for CALT. *CALICO Journal, 12(1),* 37-58.
67. OECD. (2010). *PISA Computer-based assessment of student skills in science*. http://www.oecd.org/publishing/corrigenda (accessed September 21, 2014). https://doi.org/10.1787/9789264082038-en.
68. Olsen, J. B., Maynes, D. D., Slawson, D., & Ho, K. (1989). *Comparison of paper-administered, computer-administered and computerized adaptive achievement tests*. Journal of Educational Computing Research, 5, 311-326. https://doi.org/10.2190/86RK-76WN-VAJ0-PFA3.
69. O'Malley, K. J., Kirkpatrick, R., Sherwood, W., Burdick, H. J., Hsieh, M.C. &, Sanford, E.E. (2005, April). *Comparability of a Paper Based and Computer Based Reading Test in Early Elementary Grades.* Paper presented at the AERA Division D Graduate Student Seminar, Montreal, Canada.
70. Owen, R. J. (1969). A Bayesian approach to tailored testing (Research Bulletin 69-92). Princeton NJ: Educational Testing Service.
71. Parshall, C. G. and Kromrey, J. D., (1993). *Computer testing versus paper-and-pencil: an analysis of examinee characteristics associated with mode effect.* A paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA, April (Educational Resources Document Reproduction Service (ERIC) # ED363272).

72. Pathan, M. M. (2012*). Computer Assisted Language Testing [CALT]: Advantages, Implications and Limitations*. Sebha: The University of Sebha Press.
73. Philbin, T. (2003). *The greatest inventions of all time: A ranking Past to Present* (New York: Citadel Press).
74. Pine, S.M., A.T. Church, K.A Gialluca and D. J. Weiss. (1979). *Effects of computerized adaptive testing on black and white students*. Minneapolis, MN: University of Minnesota. [Research Rep. No. 79-2].
75. Pine, S. M., & Weiss, D. J. (1987). *A comparison of the fairness of adaptive and conventional testing strategies* (Research Report 78-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program (NTIS No. AD A059436).
76. Pinsoneault, T.B., 1996. *Equivalency of computer-assisted and paper-and-pencil administered versions of the Minnesota Multiphasic Personality Inventory-2*. Computers in Human Behavior, 12, 291–300. https://doi.org/10.1016/0747-5632(96)00008-8.
77. Pintrich, P. R. (1989). The dynamic interplay of student motivation and cognition in the college classroom. In C. Ames & M. Maehr (Eds.), *Advances in motivation and achievement: Vol. 6. Motivation enhancing environments* (pp. 117-160). Greenwich, CT: JAI Press.
78. Poggio, J., Glasnapp, D., Yang, X. & Poggio, A. (2005). *A Comparative Evaluation of Score Results from Computerized and Paper & Pencil Mathematics Testing in a Large Scale State Assessment Program.* The Journal of Technology, Learning and Assessment, 3(6), 5-30.
79. Pommerich M., (2004) *Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests*. The Journal of Technology, Learning, and Assessment, 2(6) (2004).
80. Pomplun, M., Frey, S., & Becker, D. F. (2002). *The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension.* Educational and Psychological Measurement, 62(2), 337-354. https://doi.org/10.1177/0013164402062002009.
81. Powers, D. E., & O'Neill, K. (1992). Inexperienced and anxious computer users: Coping with a computer- administered test of academic skills. The Praxis Series: Professional assessments for beginning teachers. Princeton, NJ: Educational Testing Service.

82. Powers, D. E., & O'Neill, K. (1993). *Inexperienced and anxious computer users: coping with a computer administered test of academic skills.* Educational Assessment, 1, 153–173. https://doi.org/10.1207/s15326977ea0102_4.
83. Russell, M., & Haney, W. (1996). *Testing writing on computers: Results of a pilot study to compare student writing test performance via computer or via paper-and-pencil.* Retrieved July 12, 2011, from ERIC database.
84. Ryan, A. M., & Ployhart, R. E. (2000). *Applicants' perceptions of selection procedures and decisions: a critical review and agenda for the future.* Journal of Management, 26, 565–606. https://doi.org/10.1177/014920630002600308.
85. Schaeffer, G. A, Bridgeman, B., Golub-Smith, M.L., Lewis, C., Potenza, M.T., & Steffen, M. (1998). *Comparability of Paper-and-pencil and Computer Adaptive Test Scores on the GRE General Test.* (research report 98-38). Princeton, NJ: Educational Testing Service.
86. Schaeffer, G. A., Steffen, M. Golub-Smith, M. L., Mills, C. N., & Durso, R. (1995). *The introduction and comparability of the computer-adaptive GRE General Test* (Research Rep. No. 95-20). Princeton NJ: Educational Testing Service.
87. Schmidt, F. L., Urry, V. W., & Gugel, J. F. (1978). *Computer assisted tailored testing: examinee reactions and evaluations.* Educational and Psychological Measurement, 38, 265–273. https://doi.org/10.1177/001316447803800208.
88. Schmitt, N., Gilliland, S. W., Landis, R. S., & Devine, D. (1993). *Computer-based testing applied to selection of secretarial applicants.* Personnel Psychology, 46, 149–165. https://doi.org/10.1111/j.1744-6570.1993.tb00871.x.
89. Shuttleworth, M. (2009). Repeated measures design. *Experiment Resources*. http://www.experimentresources.com/repeated-measures-design.html (accessed January 25, 2012).
90. Smith, B., & Caputi, P. (2007). Cognitive interference model of computer anxiety: Implications for computer-based assessment. *Computers in Human Behavior, 23*(3), 1481-1498. https://doi.org/10.1016/j.chb.2005.07.001.
91. Smith, M. N. & Kotrlik, J. W. (1990). *Computer Anxiety Levels of Southern Region Cooperative Extension Agents.* Journal of Agricultural Education, 31(1), 12-17. https://doi.org/10.5032/jae.1990.01012.
92. Stricker, L. J., Wilder, G., & Rock, D. A. (2004). *Attitudes about computer-based test of English as a foreign language.* Computers in Human Behavior, 21 (1), 37-54. https://doi.org/10.1016/S0747-5632(03)00046-3.

93. Taylor, C., Kirsch, I., Eignor, D., & Jamieson, J. (1999). *Examining the relationship between computer familiarity and performance on computer-based language tasks.* Language Learning, 49, 219–274. https://doi.org/10.1111/0023-8333.00088.
94. Thomas, D. (2008). *The digital divide: What schools in low socioeconomic areas must teach.* The Delta Kappa Gamma Bulletin, summer 2008, 12-17.
95. Tung, P. (1986). *"Computerized Adaptive Testing: Implications for Language Test Developers."* Technology and Language Testing, edited by C. W. Stansfield. Washington, DC: TESOL.
96. Vispoel, W.P., (2000). *Computerized versus paper-and-pencil assessment of self-concept: Score comparability and respondent preferences*. Measurement and Evaluation in Counseling and Development, 33, 130–143.
97. Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). *Individual differences and test administration procedures: A comparison of fixed-item, computerized-adaptive, and self-adapted testing*. Applied Measurement in Education, 53, 53-79. https://doi.org/10.1207/s15324818ame0701_5.
98. Vispoel W. P., Wang T., & Bleiler T. (1997). *Computerized adaptive and fixed-item testing of music listening skill: A comparison of efficiency, precision, and concurrent validity.* Journal of Educational Measurement, 34, 43–63. https://doi.org/10.1111/j.1745-3984.1997.tb00506.x.
99. Wainer, H. (1990). *Introduction and history.* In H. Wainer (Ed.), Computerized adaptive testing: a primer (1-22). Hillsdale, NJ: Lawrence Earlbaum.
100. Wallace, P. E., and Clariana, R. B., (2000). *Achievement predictors for a computer-applications module delivered via the world-wide web.* Journal of Information Systems Education 11 (1) 13–18. [http://gise.org/JISE/Vol11/v11n1-2p13-18.pdf].
101. Wang, S. D., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). *A meta-analysis of testing mode effects in grade K-12 mathematics tests.* Educational and Psychological Measurement, 67(2), 219-238. https://doi.org/10.1177/0013164406288166.
102. Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2008). *Comparability of computer-based and paper-and-pencil testing in K-12 assessment: A meta-analysis of testing mode effects*. Educational and Psychological Measurement, 68, 5-24.
103. Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement, 38* (1), 19-49. https://doi.org/10.1111/j.1745-3984.2001.tb01115.x.

104. Ward, T., Hooper, S., & Hannafin, K. (1989). *The Effect of Computerized Tests on the Performance and Attitudes of College Students.* Journal of Educational Computing Research (pp. 327 - 333). https://doi.org/10.2190/4U1D-VQRM-J70D-JEQF.

105. Watson, B., (2001). *Key factors affecting conceptual gains from CAL*. British Journal of Educational Technology 32 (5) 587–593. https://doi.org/10.1111/1467-8535.00227.

106. Way, D. (2010). *Some perspectives on CAT for K-12 Assessments.* Presented at the 2010 National Conference on Student Assessment, Detroit, MI.

107. Weiss, D. J., & Kingsbury, G. G. (1984). *Application of computerized adaptive testing to educational problems.* Journal of Educational Measurement, 21, 361-375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x.

108. Wenemark, M., Persson, A., Brage, H. N., Svensson, T., & Kristenson, M. (2011). Applying motivation theory to achieve increased response rates, respondent satisfaction and data quality. *Journal of Official Statistics*, *27*(2), 393-414.

109. Wilder, G., Mackie, D., & Cooper, J. (1985). *Gender and computers: two surveys of computer-related attitudes.* Sex Roles, 13, 215–228. https://doi.org/10.1007/BF00287912.

110. Wilson, F. R., Genco, K. T., & Yager, G. G. (1985). *Assessing the equivalence of paper-and-pencil vs. computerized tests: Demonstration of a promising methodology.* Computers in Human Behavior, 1, 265–275. https://doi.org/10.1016/0747-5632(85)90017-2.

111. Wise, S. L., Barnes, L. B., Harvey, A. L., & Plake, B. S. (1989). *Effects of computer anxiety and computer experience on the computer-based achievement test performance of college students.* Applied Measurement in Education, 2, 235–241. https://doi.org/10.1207/s15324818ame0203_4.

112. Wise, S. L., & DeMars, C. E. (2003, June). *Low examinee effort in low-stakes assessment*: *Problems and potential solutions*. Paper presented at the annual meeting of the American Association of Higher Education Assessment Conference, Seattle, WA.

113. Woodrow, J. E. J. (1992). *The influence of programming training on the computer literacy and attitudes of pre-service teachers.* Journal of Research on Computing in Education, 25 (2), 200-219. https://doi.org/10.1080/08886504.1992.10782044.

114. 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. https://doi.org/10.1016/0346-251X(95)00051-K.