European Journal of English Language Teaching
ISSN: 2501-7136
ISSN-L: 2501-7136
Available on-line at: www.oapub.org/edu

doi: 10.5281/zenodo.268576 Volume 2 | Issue 1 | 2017

TRANSITIONING TO AN ALTERNATIVE ASSESSMENT: COMPUTER-BASED TESTING AND KEY FACTORS RELATED TO TESTING MODE

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

Abstract:
Computer-Based Testing (CBT) is becoming widespread due to its many identified positive merits including productive item development, flexible delivery testing mode, existence of self-selection options for test takers, immediate feedback, results management, standard setting and so on. Transitioning to CBT raised the concern over the effects of testing administration mode on test takers' scores compared to Paper-and-Pencil-Based testing. In this comparability study, we compared the effects of two different media (CBT vs. PPT) by investigating the score comparability of General English test taken by Iranian graduate students studying at Chabahar Maritime University to see whether test scores obtained from two testing modes were different. To achieve this goal, two versions of the same test were administered to 100 intermediate-level test takers organized in one testing group in two separate testing occasions. Using paired sample t-test to compare the means, the findings revealed the priority of CBT over PPT with .01 degree of difference at $p<05$. Utilizing ANOVA, the results indicated that two prior computer familiarity and attitudes external moderator factors had no significant effect on test takers' CBT scores. Furthermore, according to the results, the greatest percentage of test takers preferred test features presented on computerized version of the test.

Keywords: computer-based testing, paper-and-pencil-based testing, computer familiarity, computer attitude, test preference

[i] Correspondence: email hashemi.seyyedmorteza@gmail.com, m.hashemi@cmu.ac.ir

1. Introduction

Advances in technology have always had an impressive role in the development of human life. Sometimes technological developments have such great influences on human life that some scholars and sociologists categorize mankind history based on the produced technological tools. Technology has been greatly changing the way we live, work, think, communicate and interact with the others, and its strong continuous endless impact on all aspects of our lives is obvious (Challoner, 2009). According to the assessment researcher, Stuart Bennett who is interested in doing research in measurement, new technology's transformative impacts on assessment domain makes it possible to impel someone manage something well and satisfactorily by building some tests based on the conceptualization of preconditions and qualifications. He also declares that by utilizing technological assessment tools to create tests, test takers' performance can be practically assessed through computer based simulations, item and item bank creation and also scoring process. Besides, large-scale delivery test is made possible by using technology and computer in assessment domain (Bennett, 1999, p. 11). New types of assessment have been taken up in educational settings in USA in order to incorporate CBT into the assessment field and to help test designers develop the same test conditions as that of paper-based test for all test takers regardless of test population size (Al-Amri, 2009). Although a serious discussion on the development of Computer-Based Testing (Henceforth CBT) and a great deal of research on developing and implementing high stakes computerized version of testing program began in 70s A.D. decade by some leading works such as ASVAB (Armed Services Vocational Aptitude Battery) program done by USA Defense Department, the Graduate Record Examination (GRE), Test of English as a Foreign Language (TOEFL) and etc.), the real history of computerized fixed-length testing goes back to the decade of 30s A.D. 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 CBT which has been expedited by CAL (Computer-Assisted Learning). CBT has changed the nature of language assessment field with its potential benefits and capabilities. In fact, CBT may assist language assessment field by helping overcome many common administrative and logistical problems that are widespread in traditional fixed-length testing environment. In fact, by offering new approaches and basic advantages such as easier and more precise test scoring and reporting, item innovation, item generation, greater security, standardization, and test efficiency, test booklets and answer sheet elimination, more flexible scheduling, reduced measurement

errors, and etc., CBT opened new windows and laid foundations for future assessment in educational testing.

In examining perceptions on CBT, several issues have been identified to organize the advantages and challenges of CBT. The most important benefit of CBT is the innovation, efficiency and productivity that can be achieved in this area (Al-Amri, 2009). In CBT, input materials are presented in text, graphics, audio, and video which simulate target language situations and develop the authenticity of test tasks by enhancing the interaction between test takers and test tasks. In education, CBT is also used to administer the test to evaluate the language proficiency of English learners (Fleming & Hiple, 2004). CBT assesses test taker's language ability accurately by providing more efficient standardization of test administration conditions (Al-Amri, 2009). The same and consistent test conditions are provided by test developers in CBT (Al-Amri, 2009) and the same instructions, materials and information are presented in an enhanced consistent and uniform way to all test takers, regardless of the tests' population size, place and time of testing. Moreover, unlike paper examinations in conventional classrooms, immediate viewing of scores on screen is provided in CBT to give test takers the instant feedback. Immediate feedback, accurate test result reports and the possibility of printing the basic testing statistics are other advantages of using computer in assessment field that enable test takers take the test at any time (Mojarrad et al., 2013). CBT provides improved test security, requires less time to finish (Laurier, 1999), creates more positive attitude towards test (Madsen, 1986) and individualizes test experience.

The issue that currently needs more attention and prompt investigation of researchers is to study the testing mode effects on comparability and equivalency of the data obtained from two modes of presentation, i.e. traditional paper-and-pencil (PPT) and computerized tests. Comparability studies in second language tests are in short supply, and the importance of conducting comparability studies in local settings to detect any potential test-delivery-medium- especially when a traditional PPT is converted to a computerized one should be considered.

The critical issue of establishing comparability and equivalency of computerized test with its paper-and-pencil counterpart is of prime importance. Some research have focused on the equivalency of computer and paper-administered tests in terms of scores (Choi, Kim, and Boo, 2003; Kenyon & Malabonga, 2001; Khoshsima & Hashemi, 2017). Recently, some studies have been conducted to indicate that in order to replace computer-based 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 is not violated. But actually, there is no agreed upon

theoretical explanation for the test mode effects. The comparability is achieved through equivalent scores of two test versions.

Since in Iran, however, computerized testing is still at an early experimental stage, the present study would be conducted to provide some helpful and informative findings for those learners, teachers, testing practitioners and researchers who seek to know the possibility of replacing computerized tests with paper-pencil ones. In this study, the testing mode effects on the final performance of test takers is investigated to show whether there is any significant difference between two versions of the same test. It means that whether there is any discrepancy that violates the reliability and validity of the computerized counterpart; the computerized version that is supposed to be replaced with the conventional paper-and-pencil version of the test. In the case of Choi, Kim and Boo (2003), significant cross-mode differences in means of listening, grammar, and vocabulary subtests were examined and the largest cross mode discrepancy was observed in the reading comprehension subtest.

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 (2005) 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. 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 show that students with a good knowledge of computer use feel 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 the performance of testing mode. Some indefinite conclusions concerning to the impact of computer familiarity on performance were resulted from other studies. In one investigation, Lee (1986) distributed a computer experience questionnaire among participants and administered an arithmetic reasoning test via paper and computer medium to reach the conclusion that low- and high-computer use groups showed no significant differences in performance.

Furthermore, 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. Inevitable questions about test takers' reactions and attitudes towards computerized version of paper-and-pencil test are

raised after the introduction of 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. 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. 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).

Student preference may be considered as another factor whose relation with the performance of test takers on CBT should be examined. Some students have necessary prior familiarity and experience of using computers to play games and receiving some of their instructions through computers. Due to the possibility of customizing the assessment based on personal preferences, some people prefer to take CBT version of the test. For instance, all students have the option to select their own background color and font size preference on computer screen. Although some students may prefer CBT, others may prefer paper and pencil-based test (Cater et al., 2010; Russell et al., 2010).Some test takers prefer paper-based testing process because they are accustomed to taking notes and circling questions and/or answers for later review.

2. Literature Review

Popular computerized testing has been increasingly implemented across the world so far. Countries such as United States of America and United Kingdom have initiated use of computers in their testing and assessment environment for about three decades.

When computerized version of examinations has appeared, researchers began making comparisons between PPT and CBT. Consequently, comparability studies were conducted to study testing mode effect. Translation of paper and pencil assessment into computerized version often requires that the computerized form be comparable to its conventional paper and pencil one and the scores and the results obtained from two identical test forms approximate to each other. Interchangeability is required when students may take the same test in either mode. In fact, validity of the computerized version of a test must be confirmed by the same methods of validity determination for its traditional.

According to American Educational Research Association (AERA), in the case of using more than one way of test administration or recording the marks obtained from the test (such as marking the right answers in a booklet, separate answer sheet, or onscreen) the guidelines and instructions should express obviously that the scores

received from these ways are equivalent and interchangeable (American Educational Research Association, 1999, p. 70).

Empirical research on cross-mode comparability should be conducted to answer whether test scores are equivalent across modes in order to replace CBT with PPT. Although CBT offers some benefits over its traditional counterpart (Poggio, Glasnapp, Yang and & Poggio, 2005), comparability and equivalency of test scores between two test administration modes have been the real concerns for educators, scholars, practitioners and designers in assessment field (Lottridge, Nicewander, Schulz, & Mitzel, 2008).

Evaluating the comparability of CBT and PPT scores is critical before introducing the computerized assessment into any educational context. The main objective of a comparability study is to determine if test results obtained from two versions of the same test are equivalent. International Guidelines on Computer-Based Testing (International Test Commission, 2006) states that scores received from CBT and its conventional counterpart should be equivalent. The standards stated by International Test Commission are also supported by classical true-score test theory which is considered as the cornerstone of CBT and PPT (Allen & Yen, 1979). According to this theory, a test taker is expected to receive nearly the same test scores in two modes of test administration. The standards were examined in many comparability researches and supported by some of the empirical studies (e.g. OECD, 2010). According to Boo et al. (2012), the scores obtained from computer and paper-based tests were comparable in terms of internal consistency, criterion and construct validities, means and standard deviations. Test takers also preferred computer counterpart of conventional paper-based test and had positive attitudes towards it. Choi, Kim, and Boo (2003) reported that the results of paper and computer versions of the standardized English Language test administered to postsecondary level language learners were comparable across listening and reading comprehension, grammar and vocabulary subtests which have been proved to measure the same constructs by confirmatory factor analysis. Of course, a more comprehensive and detailed investigation of all these subtests indicated that the reading comprehension and grammar subtests revealed weakest and strongest comparability, respectively (p. 316). In a last comparability study, Khoshsima & Hashemi (2017) concluded that test scores of test takers did not vary in both PPT and CBT. Their findings confirmed the equivalency of test takers' scores obtained from two different testing modes.

Florida Department of Education (2006) reported that early examinations of the relationship between computer familiarity and test performance showed significant difference. It means empirical evidences confirmed lower scores of test takers who had

less experience and familiarity with computers. But it also asserted that recent studies show no relationship between them (Florida Department of Education, 2006). In another research, no relationship was found between prior computer experience and computerized TOEFL test performance (Taylor et al., 1999). 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 more examination on prior frequent use of computer as a moderator variable in CBT have to be considered.

Attitudes towards computerized test play a crucial role in implementing CBT successfully. Attitudes towards computer can be influenced by some other contextual factors such as age, gender, socioeconomic status and etc. Although prior attitudes towards computers may have a direct relationship with prior computer experience, these two constructs are completely distinct from each other. According to Eagly and Shelly (1998), attitude is positive or negative feelings towards a psychological object. In another definition of attitude, Loyd and Gressard (1985) name four components including anxiety, confidence, liking, and usefulness that organize attitude towards computer. Al-Amri (2009) utilized some special sections of CAS questionnaire to study learners' attitudes towards computer use. In spite of the fact that students show high preference for CBT, his research findings indicate no relationship between learners' attitudes and their performance on CBT. The same study has been done by Youdbakan and Uzunkavak on learners' attitudes towards computer and CBT in both private and state schools. A researcher made attitude scale was distributed among 784 Turkish primary school learners who participated in the study. The data that was collected from the piloted researcher made questionnaire indicated no significant difference in attitudes towards computer. But the students of state schools showed more positive attitudes towards CBT. Generally, no association effect was found between attitudes towards CBT (Youdabakan and Uzunkavak, 2012).

In addition to computer familiarity and computer attitude, testing mode preference of test takers that is typically related to high stakes standardized test administration has attracted much attention in recent researches. Like this study, many studies have been done to examine the preference of test takers on testing administration mode (Al-Amri, 2009; Flowers et al., 2011; Higgins et al. 2005; Khoshsima et al., 2017; Yurdabakan & Uzunkavak, 2012). Testing mode preference is a contributing factor that should be considered in comparability studies. In a research conducted by Flowers et al. (2011), there was a high preference for CBT, and test takers' preference had negative correlation with test takers performance on CBT. According to their findings, although test takers show high preference for taking CBT, they

outperformed on PPT. According to Al-Amri (2009), although test takers preferred to take CBT, their test performance was better on PPT. His research findings show no relationship between test performance and testing mode preference. In another similar study, no correlation between testing mode preference and testing performance of test takers was found (Khoshsima & Hashemi, 2017).

Since evaluating the comparability of paper-based and computer-based tests is crucial before introducing computer aided assessment into any context, the present study, first, seeks to examine cross-mode effects on test takers' General English scores. The second purpose of the study is to examine the relationship of computer familiarity, prior attitudes towards computer and testing mode preference with testing performance on CBT version. Considering both theoretical and pedagogical perspectives, the following questions are addressed in this study to accomplish the main purposes:

RQ1. Is there any statistically significant difference between computer-based language testing and paper and pencil-based one when assessing General English of Iranian graduate students?

RQ3. Is there any relationship between two computer familiarity and prior attitudes towards computer external variables and Iranian graduate students' testing performance on CBT version of the test?

RQ3. Do participants' prior testing mode preferences affect their performance on CBT?

3. Methodology

3.1 Research design

The present research that covered both comparison and correlational studies explored the comparability of paper and computer-based testing in a General English context and the correlation between some external moderator factors including test takers' characteristics such as computer attitude, prior computer experience and testing mode preferences that were believed to be meaningfully related (Warner, 2013) to their testing performance on computer-based language testing in comparison with paper-based version. In order to reach more solid conclusions in this research, a mixed-method approach including both qualitative and quantitative instruments were utilized to investigate the difference between test results due to its advantages such as easy and fast data collection, consistency and accuracy of collected data and proper descriptive and inferential results. The mixed-methods approach of the study combined multiple

choice achievement tests, questionnaires and interviews that were employed in this study.

3.2 Participants

The selected participants for the present study were 100 graduate students of Maritime University of Chabahar. After administering New Interchange placement test to 186 graduate students to identify intermediate level students, 128 homogenous students were selected. 28 participants were removed because they were unwilling or unable to complete the study. Of the remaining total participants who were assigned to one testing group to take two versions of the same test, there were slightly more girls (n=60%) than boys (n=40%). The age range of all the 100 students who had signed the consent form to participate in the study was between 23 to 28 years. And, the mean age was 24.5 (Table 1).

Table 1: Gender frequency distribution

	Testing group one		**Testing group two**	
Gender	**frequency**	**percentage**	**frequency**	**percentage**
Male	22	44	18	36
Female	28	56	32	64
Total	50	100	50	100

3.3 Instruments

New-Interchange Placement Test was implemented to the participants of the study to the purpose of checking their homogeneity and to make sure that they are homogeneous in terms of general English knowledge and proficiency. The testing group took two versions of a test derived from General English book on separate testing sessions with four weeks interval. The four weeks interval was to mitigate the practical potential, fatigue effects and testing effects. The study employed General English multiple-choice achievement test as the main research data instrument to compare the mean of scores received from both testing modes. The paper version of the test was converted into computer version using ClassMarker.com website.

Unlike the paper-based format in which all the question items were presented in three pages, with CBT version of the test, test takers were presented one question per screen. When the question item was presented to the test taker, s/he should click on the letter of the right answer and then proceeded to the next item. Like PPT, test takers could review previously answered questions and change them due to the nature of this kind of computerized testing.

The items order was the same in both versions of the test. To examine the internal consistency (Cronbach's alpha) of the test on each testing mode, the responses of testing group of the present study were investigated and relatively high reliability coefficients (α = .865) and (α =.880) for PPT and CBT, respectively, were achieved.

The second procedure that was employed in this research attempted to answer the research question two. It was used to see if there was any relationship between two computer familiarity and prior computer attitudes external moderator variables and test takers' testing performance on CBT. To meet this objective, the standard Loyd Gressard Computer Attitude Scale (Loyd and Gressard, 1985) that was validated by Berberoglu and Calikoglu (1992) was distributed to the test takers after implementing CBT version of the test. It should be mentioned that high reliability coefficient was reported on the total score by Loyd and Gressard (1985). Christensen and Knezek (1996) also reported high reliability coefficient value of .95 and stable factorial validity. After examining the internal consistency of CAS questionnaire distributed to the participants, fair reliability coefficient value of .84 was obtained for this study.

Another instrument to collect the research data concerning to the third research question was a simple question mentioned at the bottom of exam paper and screen, i.e. *would you prefer taking test on paper – no difference – computer* to examine the relationship between testing mode preference and performance. Due to the importance of relationship between testing mode preference and testing performance when conducting PPT and CBT, our third research question examined the correlation between test takers' testing mode preference and their performance on either testing mode.

The last qualitative instrument was a formal semi-structured interview through which a series of data was collected and coded to be analyzed quantitatively. The qualitative research data that was collected to support the quantitative research data came from conducting semi-structured interviews with 30 participants who were randomly selected from the testing group. Based on the previous literature, the questions of the interview were developed by the researcher and then content analyzed by two instructors of TEFL in CMU.

3.4 Procedure

New-Interchange Placement Test was administered to 186 graduate students to the purpose of checking their homogeneity. Consequently, the intermediate level students were selected to participate in the research. Then, the testing group was given both versions of General English multiple-choice achievement test in two separate testing sessions with four weeks interval. At the end of both exams, testing group answered the simple question *would you prefer taking the test on paper – no difference – computer* to

explore the relationship between testing mode preference and testing performance. Before taking CBT version, test takers received a simple sample computerized task and oral instruction on how to take the computerized version of the test. After becoming familiar with the CBT environment, every test taker was given a unique registration code to register into the assigned group created in the website. Test takers had 40 minutes to answer 50 question items (the time given to complete the sample exercise before administration of CBT was not included). On the onscreen test, students received one question per screen. Students clicked on the letter of the correct answer choice and then proceeded to the next question. Like paper-based testing, students could go back, review and change previously answered questions in CBT.

And T the last stage, formal semi-structured interviews were conducted through which a series of related qualitative data was collected and coded to be analyzed quantitatively. The participants were asked about their attitudes towards the features of two modes of testing administration, testing mode preference, development of positive or even negative attitudes and their reasons for possible changing mode preference. Some of the participants who changed their preference were also asked about their reasons to change their preferences after taking CBT. In the focus group semi-structured interview, the participants were asked a series of pre-determined open-ended questions on the issue based on a list of topics in a particular order (Interview Guide). The researcher used the interview guide printed on paper that was required to be observed during the conversations in order not to stray from the interview procedure. The interview for each participant took about 7-10 minutes. Totally, 30 interviews took about 250 minutes in one session. The components of interview were a brief introduction of CBT and its history, some questions about participants' testing mode preference, and their comments about CBT and PPT features.

4. Results and Discussions

The usual procedures for comparability are psychometric characteristics such as the distribution, rank, and correlation of scores on two tests (Choi et al., 2003), shape of the score distribution, reliability, and conditional standard error of measurement (Wang & Kolen, 2001). Aforementioned criteria that are usually considered in comparability study of CBT and PPT are compatible with the criteria that are declared by some testing organizations such as the International Test Commission (ITC) and the American Psychological Association (APA). ITC testing organization states that the designers of computerized tests should produce the interchangeable scores whose means and standard deviations are the same as their PPT counterparts (International Test

Commission, 2006, p. 156-157). A majority of research conducted on PPT and CBT comparison focused on the differences in means and standard deviations (e.g. Khoshsima et al., 2017; Makiney, Rosen, & Davis, 2003; Pinsoneault, 1996). Before exploring the comparability of paper and computer-based testing in the General English context by employing paired sample t-test test, we examined the normality of the data distribution.

Shapiro-Wilks and Kolmogorov-Smirnov statistical tests were used to provide objective judgement of data distribution normality. Anyway, the result of normality testing is displayed in Table 2 statistically.

Table 2: Normality distribution test

Tests of Normality						
	Kolmogorov-Smirnov[a]			**Shapiro-Wilk**		
	Statistic	**D.F.**	**Sig.**	**Statistic**	**D.F.**	**Sig.**
PPT	.115	100	**.890**	.931	100	**.912**
CBT	.165	100	**.868**	.954	100	**.951**

From Table 2, it was concluded that the research data obtained from two PPT and CBT versions of General English tests administered to testing groups of graduate students in two separate testing sessions were normally distributed.

We continued data analysis by conducting paired t-test. The main goal of t-test series conducted in this section was to examine if there was any statistically significant difference in participants' testing performance in PPT and CBT. According to the results, the mean score of test takers on PPT testing performance (M = 2.48, SD = .16135) was lower than the mean score of test takers on CBT testing performance (M = 2.51, SD = .15982) (Table 3). Then, of the two versions of the test taken by testing group, the highest mean score was found for the performance of testing group on CBT. Furthermore, the higher standard deviation for PPT results indicated that the dispersion of scores from mean score for CBT was lower.

Table 3: Descriptive Statistics of test scores in both PPT & CBT

Groups		**Independent Samples Statistics**			
		Mean	N	Std. Deviation	Std. Error Mean
General English Test	PPT	**2.4815**	100	**.16135**	.03608
	CBT	**2.5155**	100	**.15982**	.03574

Then, according to the inferential analysis, there was a statistically significant difference between test takers' mean scores from PPT (M = 2.48, SD = .16) and CBT (M = 2.51, SD =

.15); (t (98) =-4.773, P=0.000) (Table 4). It can be concluded that there is a statistically significant difference between the mean scores of the graduate students on both PPT and CBT versions of the test.

Results of paired t-test comparing mean scores of test modes are indicated in Table 4. The aim of this test was to gather further evidence to ensure whether two testing administration modes were showing interchangeable results. According to Table 4, t-test revealed that the t-statistic value was 0.000 with 29 degree of freedom at P<0.05. The corresponding two-tailed p-value was 0.000 that was smaller than 0.05.

Table 4: Paired t-test results for both PPT and CBT modes of administration

	Paired Differences					t	D.F.	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
PPT – CBT	-.03400	.03185	.00712	-.04891	-.01909	-4.773	98	**.000**

In order to answer the second research question, ANOVA statistical test was to examine the significant difference between computer familiarity and attitudes and testing performance of students. The results in Table 5 indicate that the F Observed value for the students' prior computer familiarity and CBT is 1.82 (P =0.895 > 0.05). Based on these results, it can be concluded that the students' computer familiarity does not have any significant correlation or interactive effect between computer familiarity and on CBT performance.

Additionally, the F observed value for the effect of the prior attitudes towards computer on CBT performance is 1.87 (P = .456 > 0.05). Therefore, it can also be concluded that the prior computer attitudes does not have any significant influence on CBT performance of test takers. Based on the findings, no significant correlation was seen between the participants' attitudes towards computer and CBT performance.

Table 5: ANOVA results of interactive effect of computer familiarity and attitudes on CBT performance

Source	DF		F	Sig.
Mode*computer familiarity scale	Sphericity Assumed	1	1.82	.895
Mode*computer familiarity scale	Sphericity Assumed	1	1.87	.456

To answer the research question three, the relationship between testing mode preference and testing performance was examined. To reach this aim, the correlation

between participants' responses to the simple question appearing at the end of PPT exam, i.e. *would you prefer taking test: 1.On paper 2.No difference 3.On computer* and their mean scores obtained from CBT version of the test was examined. The answers that participants gave to the question were coded as 1, 2 and 3 for "On paper", "No difference", and "On computer". Table 6 and 7 display the results of the correlations between pre and post-CBT testing mode preferences and CBT testing performance variables.

Table 6: Correlations of pre-CBT testing mode preference and mean of CBT scores

		Mean of CBT
Pre-CBT testing mode preference	**Pearson Correlation**	.142
	Sig. (2-tailed)	**.312**
	N	100

The Pearson product-moment correlation was run to examine the relationship between pre-CBT testing mode preference and testing performance. According to the results, for the testing group, the answers of participants to the first testing mode preference question (M=1.86, SD=.89) and CBT performance (M=2.48, SD=.161) were not significantly correlated; .142(98) =.312, P >1. According to the findings it can be concluded that pre-CBT testing preference mode is not correlated with test takers' scores in CBT.

Table 7: Correlations of post-CBT testing mode preference and mean of CBT scores

		Mean of CBT
Post-CBT testing mode preference	**Pearson Correlation**	.192
	Sig. (2-tailed)	**.436**
	N	100

The Pearson product-moment correlation was also run to examine the relationship between post-CBT testing mode preference and CBT testing performance. According to the results, for the testing group, the answers of participants to the second testing mode preference question (M=2.46, SD=.81) and their CBT performance (M=2.51, SD=.159) were not significantly correlated; .192(98) =.436, P>1.

In the next stage, we examined if test takers performed better on their preferred testing mode according to their pre and post-CBT testing mode preference and testing performance. The descriptive statistics are shown in Table 8.

Table 8: Descriptive statistics of testing group's performance according to participants' pre and post-CBT preference and testing performance in two testing administration modes

Testing sessions	Preferred testing mode	N	Mean		Std. Deviation	
			Pre-CBT	Post-CBT	Pre-CBT	Post-CBT
PPT	**Paper**	55	36.12	**38.19**	10.74	17.20
	No difference	15	39	**45.12**	6.77	10.93
	Onscreen	30	48.76	**42**	25.93	15.94
CBT	**Paper**	15	46	**58**	12.52	49.33
	No difference	10	48	**48**	13.85	13.85
	Onscreen	75	56	**44.35**	16.87	16.87

According to the findings, in paper-based testing session, participants who preferred to take paper-based version of the test outperformed on CBT (M=38.19) and those who preferred to take computerized version of the test performed in PPT (M=48.76). After implementing CBT version of the test, the answers of testing mode preference question appeared at the bottom of the screen was analyzed. As it was shown in Table 8, those participants of computer-based testing session who preferred to take PPT version of the test performed better on CBT (M=58) and those who preferred to take the test on CBT performed better on PPT (M=56). The findings indicated that there was no interaction between testing mode preference and testing performance of participants. Then, it can be concluded that testing mode preference does not affect test validity.

The qualitative research data that was collected to support the quantitative research data came from conducting a semi-structured interview with 30 participants who were randomly selected from two testing groups. In interview session, if the participant had changed his/her testing mode preference after taking the CBT, s/he would have been asked about her/his reasons to change the preference. To analyze the qualitative data, the interview conversations were transcribed. In transcription, just the relevant sections of recorded conversations were picked up. Once transcription of the data has been completed, content analysis was conducted on transcribed data by identifying all the main concepts. The content analysis involved a thematic analysis of the received data. In thematic analysis, similar statements and responses to the same question were coded and categorized under a common theme (Seidman, 1998). The main relevant and meaningful notions and concepts were identified and categorized under common themes.

5. Conclusion

The purpose of the current study was to investigate the equivalency of test results in CBT and PPT by comparing the test results of two modes of testing administration among graduate students of Chabahar Maritime University in Iran. Moreover, it sought to probe the probable relationship of prior computer familiarity, attitude towards the use of computer and testing mode preference with testing performance on CBT. Therefore, this study employed a quantitative design to determine whether there was any difference between test scores on PPT and CBT as well as finding out any relationship between aforementioned moderator factors and their test results on CBT. It also enjoyed a qualitative design using focus group interview to find out what was the preference of test takers in test modes and their justifications for their preferences.

For the first research question which aimed at investigating the comparability of scores obtained through two PPT and CBT versions of the test, paired t-test was conducted. It was used to compare the means of two sets of scores of testing group obtained in two different testing sessions. Based on the findings, it was concluded that there was statistically significant difference in the mean scores of testing group in two testing sessions as a whole (p=.000). The findings of the research question one were compatible with the results of (Coniam, 2006; Fulcher, 1999) who claim that assessments are not comparable across modes.

In comparability studies on CBT and PPT, it is important to take into account the factors influencing the results on computerized tests especially when there is a significant or even slight difference between test scores. Some of these influencing external variables that have been investigated by many researchers due to increasing development and changing the interest in using computers are computer familiarity and attitude towards the use of computers. This is why in this study; the second main question was examining the relationship between these variables and test performance on CBT. If there was any relationship, the difference between two test modes could be attributed to the influence of these constructs irrelevant variable on CBT result.

The findings revealed that there was no interactive effect of computer attitudes and computer familiarity variables with testing performance of participants on CBT. It means that whether test takers have high or low degrees of prior positive or negative attitudes towards computer and computer familiarity, there is not any advantage or disadvantage while performing on CBT. Additionally, it supports the construct validity of CBT as this construct-irrelevant variable is not considered as a component or part of the construct that is measured by CBT version of the test.

Moreover, the overall descriptive statistics of prior testing mode preference and testing performance of different preference groups' analysis answered negatively the research question 3. These findings indicated that there was no necessarily positive interaction between testing mode preference and testing performance. The reason might be the novelty of CBT in the target setting. The findings of the present study were in consistent with the result of Khoshsima et al.'s (2017) study that found out test takers with positive attitudes towards the use of computer did not perform better on CBT. Testing mode preference of test takers of testing group was examined before and after exposure to CBT. Then, the testing mode preference was categorized under two pre-CBT and Post-CBT testing mode preferences. By analyzing two pre and post-CBT questionnaires of testing group one to study possible testing mode preference change, it was revealed that only 15% of the test takers still preferred PPT version of the test while just 10% didn't mind taking the test on either mode. The greater percentage 75% was the test takers who opted for computer as their preferred mode of testing. We concluded that the number of participants who preferred PPT and who didn't mind taking the test in either mode have changed in favor of the test takers who chose On Computer as their preferred testing mode preference.

References

1. Al-Amir, S., (2009).*Computer-based testing vs. paper-based testing: establishing the comparability of reading tests through the evolution of a new comparability model in a Saudi EFL context.* Unpublished doctoral dissertation. University of Essex, England.
2. Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole.
3. American Educational Research Association. (1999). *Standards for Educational and Psychological Testing*. Washington: American Educational Research Association.
4. Bennett, R. E. (1999). *How the Internet will help large-scale assessment reinvents itself*. Education Policy Analysis Archives, 9(5), 1-25.
5. Berberoglu, G. & Calikoglu, G. (1992).*The construction of a Turkish computer attitude scale.* Studies in Educational Evaluation, 24 (2), 841-845.
6. Boo, J. & Vispoel, W. (2012). *Computer versus paper-and-pencil assessment of educational development: A comparison of psychometric features and examinee preferences.* Psychological Reports, 111, 443-460.

7. Cater, K., Rose, D., Thille, C., & Shaffer, D. (2010, June). *Innovations in the classroom.* Presentation at the Council of Chief State School Officers (CCSSO) National Conference on Student Assessment, Detroit MI.
8. Challoner, J. (2009). *1001 Inventions that changed the world* (Cassell Illustrated: 2009).
9. Choi, I. C., Kim, K. S., & Boo, J. (2003*). Comparability of a paper-based language test and a computer-based language test*. Language Testing, 20, 295–320.
10. Christensen, R. and Knezek, G. (1996). *Constructing the Teachers' Attitudes toward Computers (TAC) questionnaire*. ERIC Document Reproduction Service No. ED398244.
11. Coniam, D. (2006). *Evaluating computer-based and paper-based versions of an English language listening tes*t. *ReCALL,* 18, 193-211.
12. Eagly, A. H., & Shelly C., (1998). —*Attitude Structure and Function."* In Handbook of Social Psychology, ed. D.T. Gilbert, Susan T. Fisk, and G. Lindsey, 269–322. New York: McGowan-Hill.
13. Fleming, S. & Hiple, D. (2004). *Foreign language distance education at the University of Hawai'i*. In C. A. Spreen, (Ed.), new technologies and language learning: issues and options (Tech. Rep. No.25) (pp. 13-54). Honolulu, HI: University of Hawai'i, Second Language Teaching & Curriculum Center.
14. Florida Department of Education. (2006, September 4). *What do we know about choosing to take a high-stakes test on a computer*? Retrieved May 15, 2010, from: http://www.fldoe.org/asp/k12memo/pdf/WhatDoWeKnowAboutChoosingToTake AHighStakesTestOnAComputer.pdf.
15. Flowers, C., Do-Hong, K., Lewis, P., & Davis, V. C. (2011). *A comparison of computer-based testing and pencil-and-paper testing for students with a read- aloud accommodation.* Journal of Special Education Technology, 26(1), 1-12.
16. Fulcher, G. (1999). *Computerizing an English language placement test*. ELT Journal, 53(4), 289-299.
17. Higgins, J., Russell, M., & Hoffmann, T. (2005). *Examining the effect of computer-based passage presentation on reading test performance.* Journal of Technology, Learning, and Assessment, 3(4). Retrieved July 5, 2005, from http://www.jtla.org.
18. International Test Commission. (2006*). International guidelines on computer-based and Internet delivered testing.* International Journal of Testing, 6, 143–171.
19. 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.

20. Khoshsima, H. & Hashemi, 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(2017).
21. Laurier, M. (1999). *The development of an adaptive test for placement in French.* In M. Chalhoub-Deville (ed.), *Development and research in computer adaptive language testing* (pp. 122-35). Cambridge: University of Cambridge Examinations Syndicate/Cambridge University Press.
22. Lee, J., Moreno, K. E., & Sympson, J. B. (1986). *The effects of mode of test administration on test performance.* Educational and Psychological Measurement, 46, 467-473.
23. Lottridge, S., Nicewander, A., Schulz, M. & Mitzel, H. (2008). *Comparability of Paper-based and Computer-based Tests: A Review of the Methodology.* Pacific Metrics Corporation 585 Cannery Row, Suite 201 Monterey, California 93940.
24. Loyd, B. H, & Gressard, C. (1985). *The Reliability and Validity of an Instrument for the Assessment of Computer Attitudes.* Educational and Psychological Measurement, 45(4), 903- 908.
25. Madsen, H. S & Larson J. W. (1986). *Computerized Rasch Analysis of item bias in ESL Tests.* In C. W. Stansfield (Ed.), Technology and language testing. A collection of papers from the annual colloquium on language testing research. Princeton, New Jersey.
26. Makiney, J.D., Rosen, C., Davis, B.W., Tinios, K. & Young, P. (2003). *Examining the measurement equivalence of paper and computerized job analyses scales.* Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.
27. Mojarrad, H, Hemmati, F, Jafari Gohar, M, & Sadeghi , A. (2013). *Computer-based assessment (CBA) vs. Paper/pencil-based assessment (PPBA): An investigation into the performance and attitude of Iranian EFL learners' reading comprehension.* International Journal of Language Learning and Applied Linguistics World, 4(4), 418-428.
28. OECD. (2010). *PISA Computer-based assessment of student skills in science*. http://www.oecd.org/publishing/corrigenda (accessed September 21, 2014).
29. 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.

30. 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.
31. 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.
32. Russell, M., Almond, P., Higgins, J., Clarke-Midura, J., Johnstone, C., Bechard, S., & Fedorchak, G. (2010, June). *Technology enabled assessments: Examining the potential for universal access and better measurement in achievement.* Presentation at the Council of Chief State School Officers (CCSSO) National Conference on Student Assessment, Detroit MN.
33. 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.
34. Seidman, I. (1998). *Interviewing as qualitative research: A guide for researchers in education and the social sciences* (2nd ed.). New York: Teachers College Press.
35. 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.
36. Wallace, P., & Clariana, R. (2005). *Perception versus reality – Determining business students' computer literacy skills and need for instruction in information concepts and technology,* Journal of Information Technology Education, 4, 141-151. Retrieved March 26, 2008 from http://jite.org/documents/Vol4/v4p141-151Wallace59.pdf
37. Wang, T., & Kolen, M. J. (2001). *Evaluating comparability in computerized adaptive testing: Issues, criteria and an example*. Journal of Educational Measurement, 38, 19–49.
38. Warner, R. M. (2013). *Applied Statistics: From Bivariate through Multivariate Techniques*. (2th Ed.). SUA: SAGE Publication Inc.
39. Watson, B., (2001). *Key factors affecting conceptual gains from CAL*. British Journal of Educational Technology 32 (5) 587–593.
40. Yurdabakan, I., & Uzunkavak, C. (2012). *Primary school students 'attitudes towards computer based testing and assessment in turkey.* Turkish Online Journal of Distance Education, 13(3), 177-188.