International Journal of Computer (IJC)

ISSN 2307-4523 (Print & Online)

http://ijcjournal.org/

Comparability of Computer-Based Testing and Paper-Based Testing: Testing Mode Effect, Testing Mode Order, Computer Attitudes and Testing Mode preference

Hooshang Khoshsima[a], Seyyed Morteza Hashemi Toroujeni[b*]

[a]*PhD, Associate Professor, English Language Department Faculty of Management and Humanities Chabahar Maritime University, Iran, +98 9121097812*

[b]*M.A. in TEFL, English Language Department Faculty of Management and Humanities Chabahar Maritime University, Iran, +98 9112577241*

[a]*Email: Khoshsima2002@yahoo.com*

[b]*Email: Hashemi.seyyedmorteza@gmail.com, M.hashemi@cmu.ac.ir*

Abstract

With promulgation of computer technology in educational testing, computerized testing (henceforth CBT) as green computing strategy is gaining popularity due to its advantages such as effective administration, flexible scheduling and immediate feedback over its conventional paper-based testing (henceforth PBT). Since some testing programs have begun to offer both versions of a test simultaneously, the effectiveness of CBT is queried by some scholars. Regarding to this aim, this study investigated the score equivalency of a test taken by 228 Iranian undergraduate students studying at a state university located in Chabahar region of Iran to see whether scores of two administrations of testing mode were equivalent. Then, two versions of the test were administered to the participants of two testing groups on four testing occasions in a counter balanced administration sequence with four weeks interval. One-Way ANOVA and Pearson Correlation tests were used to compare the mean scores and to find the relationship of testing order, computer attitudes and testing mode preference with testing performance. Findings of the study revealed that the scores of test takers were not different in both modes and the moderator variables were not considered external factors that might affect students' performance on CBT.

Keywords: Computer-Based Testing; Paper-Based Testing; Testing Mode Effect.

---

* Corresponding author.

1. Introduction

Technology has been greatly influencing 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 [1]. Using various assessment delivery media such as paper and pencil and computer resulted in different computer-based (henceforth CBT) and paper-based (henceforth PBT) assessment modes in recent years.

Among the guidelines published by several organizations, [2,3] devoted their standards and guidelines to CBT exclusively. The specific goal of all these guidelines issued by various professional testing organizations is to guide all the people involved in testing domain such as test teachers, test constructors, publishers and test takers and users to consider the maximum comparability and equivalency between two different modes of testing administration. The standards of International Guidelines on Computer-based Testing [2] that are supported by experimental investigations [4] declared that the equivalency of scores obtained from CBT and its conventional paper-based testing counterpart should be established to call the CBT version valid and reliable.

2. Literature review

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. In the past, limited availability and high costs of computer and the related technological tools restricted computer-based tests administration. But nowadays, the condition is reversed. In fact, by offering new approaches and basic advantages, CBT as the most accurate way opened new windows and laid foundations for future assessment in educational testing to evaluate language proficiency of English learners [5]. In fact, technology developments and widespread accessibility to computer, especially in educational contexts, have greatly influenced many areas of interests and subjects such as English testing domain [6,7]. And, this is the reason that some international macro-organizations dealing with conducting TOEFL, IELTS, GRE tests and etc. started to give their offline or online examinations in computerized version.

Peat and Franklin believe that the use of formative and summative computer based assessment leads to important advantages and benefits for both staff and students [8]. Staffs are engaged more in interacting and communicating with students and consequently students enjoy opportunities to gain extensive and immediate feedback at the time that test is terminated. However, since the aforementioned testing organizations started to administer CBT in their assessment system along with conventional PBT assessment system and several institutes and universities started to replace CBT with its conventional counterpart due to its advantages over PBT [9], the critical issue of equivalency of scores received from two testing mode administrations should be considered [10,11,12] cautiously. In spite of the fact that a transition from conventional PBT towards CBT version of assessment is ongoing [13] in testing domain, the effect of changing mode of administration from paper to screen on students' performance has not been fully investigated yet.

In this study, the testing mode effects on the final performance of test takers will be investigated to show

whether there is any significant difference between two test versions. It means that whether there is any discrepancy that violates the reliability and validity of the computerized counterpart. In some comparability studies conducted to synthesize administration mode effects of CBT and PBT [14], comparison of two sets of test scores across various testing modes leaded to contradictory results [15]. The results of some studies indicate that students outperformed in CBT versus PBT administration [16], but there was no statistically significant difference in students' performance across modes [17,18]. The results of these studies have substantially influenced current approaches to investigate comparability between two versions of a test. Some other studies have found lower scores on CBTs compared with PBTs [19,20], higher scores on CBTs compared with PBTs [21,22,23], better performance in PBT rather than CBT [24] or no test mode effects at all [25]. Although obtained findings are not entirely conclusive, there seems to be a trend indicating that the two versions are comparable across the administration mode [26,27,12]. Different hypotheses have been advanced by the researchers of the testing domain to explain such effects of testing mode administration. For examples, time limits of testing, test difficulty, cognitive processes required by test, and presence or absence of test administrator are influencing factors causing test mode effects [28].

In addition to exploring testing mode effect on equivalency or comparability of scores from different test versions, testing order effect on students' testing performance is considered as a crucial issue. Testing order effect occurs if there is a statistically significant difference in the comparison between two CBT versions of the test administered to test takers of two testing groups who take different versions in a counter-balanced administration sequence.

Furthermore, Fulcher suggested that not only the issue of equivalency of scores is important in replacing CBT with PBT, but also other equating issues such as attitudes towards the use of computers as well as testing mode preference are crucial to consider [29]. Leeson classified key factors associated with onscreen testing mode that can lead to some difficulties in implementing CBT under two main titles including factors related to the "users" and "technology used'. He declares that some variables such as gender, IT skills and ability to process information, and the level of users' aversion towards the use of computer in testing may have an influence on the performance of users on CBT [30].

Since evaluating the comparability of paper-based and computer-based tests is crucial before introducing computer aided testing into any context, based on the consideration of all the issues discussed above, the purpose of conducting this study is to assess vocabulary knowledge of undergraduate ESP students of Chabahar Maritime University (CMU), Iran, by comparing two delivery media of Vocabulary in Use test (paper vs. computer). First, in order to seek the purposes that are pivotal to this study, the present study explores the comparability of paper and computer-based testing in an English language vocabulary context and the relationship between testing order and two test takers' characteristics including their prior attitudes towards computer and testing mode preferences with their testing performance on CBT in comparison with its conventional counterpart i.e. paper-based version. The results of aforementioned research and the fact that the current study is the pioneer of large-scale comparability studies in state universities of Iran to assess language knowledge and proficiency of the ESP students on CBT have motivated the researcher to carry out the current study. Taking into consideration all of the issues discussed above in order to identify reliability of CBT versus

PBT and testing mode effects between CBT and PBT, to examine testing mode order effect on students' testing performance, and also to explore the interaction of testing mode preference and prior computer attitudes with testing performance, the following research questions have been developed accordingly:

RQ1: Is there any statistically significant difference between computer-based language testing and paper and pencil-based one when assessing vocabulary skill of the undergraduate university students of ESP courses?

1.1 Does testing order influence test takers' performance?

RQ2: Do test takers' prior computer attitudes affect their performance on CBT?

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

3. Methodology

3.1 Participants

The participants were students who were admitted to the CMU and enrolled in different ESP courses (as part of their official curriculum). The participants who attended in the main investigation were 228 homogenous English language learners. Among those 350 undergraduate students who had taken New-Interchange placement test, the homogeneity of 256 intermediate students was specified. 28 students were removed because they were unwilling or unable to complete the study. Of the remaining 228 homogeneous ones, there were slightly more men (n=56.15%) than women (n=43.85%). The average age of the students whose age ranged from 19 to 23 (M=60.25, SD=1.24) was 20.2. The 228 homogenous students who had signed the consent form to participate in our study were randomly selected and organized into two testing groups to take both PBT and CBT formats of the same test in a counter-balance administration order in four testing occasions.

3.2 Instruments

After implementing placement test to determine the homogeneity of test takers, fixed length linear paper and pencil version of English Vocabulary in Use Pre-intermediate and Intermediate Level Test was administered immediately at the end of the ESP course teaching period. This type of the test was the traditionally common form of testing in the university that all participants of this study were clearly familiar with it.

Unlike the paper-based format in which all the question items were presented in three pages, with the CBT, 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 PBT, 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 make test takers familiar with CBT environment, a simple sample computerized test consisted of 5 questions as well as oral explanations of the researcher on how to activate test taker account and how to take the computerized test were given to the participants. It is worth

mentioning that to examine the internal consistency (Cronbach's alpha) of each testing mode of two testing groups, the responses of two testing groups of the present study were investigated and high reliability coefficients (α = .906) and (α =.913) for PBT and CBT versions of testing group one, and excellent Cronbach's alpha coefficients (α = .923) and (α =.93) for CBT and PBT version of testing group two were obtained to ensure reliable tests would be used in the research

The second procedure that was employed in this research attempted to answer the research question two. It was used to see if there existed any relationship between prior computer attitudes and testing performance. To meet this objective, the standard Loyd Gressard Computer Attitude Scale [31,32] that was validated by Berberoglu and Calikoglu in 1992 was distributed to the test takers of two testing groups after implementing CBT version of the test to get their feedback on the current issues [33]. It means that the questionnaire was delivered to testing group one and testing group two in the second and first testing sessions, respectively. CAS questionnaire is an instrument that measures various subscales and aspects of attitudes towards computer. In fact, the total score of all subscales of CAS including 40 questions with choices measured on a 4 point Likert scale is used to measure computer attitude. Therefore, the values of scores that can be achieved range from a low 40 to a high of 160. For our research, higher score values indicate more positive attitudes towards computer and lower scores represent negative computer attitudes. This paper reports the total scores of CAS to measure attitudes of test takers towards computer. It should be mentioned that high reliability was reported on the total score by Loyd and Gressard [31,32]. Christensen and Knezek also reported high reliability coefficient value of .95 and stable factorial validity. It should be mentioned that after examining the external consistency of the CAS questionnaire distributed to the participants, fair reliability coefficient value of .84 was obtained for this study [34].

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 which mode? 1. on paper 2. no difference 3. on computer*, to examine the relationship between testing mode preference and testing performance. In fact, testing mode preference variable was examined before and after test takers were exposed to CBT.

The last qualitative instrument was a formal semi-structured interview through which a series of related qualitative 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 40 participants who were randomly selected from two testing groups. Based on the previous literature, the questions of the interview were developed by the researcher and then content analyzed by two experts of TEFL.

3.3 Procedure

The counter-balanced administration sequence was applied in this study because 114 homogenous students who were similar to the first testing group's students were assigned to another testing group based on the common person design to investigate the effect of testing order alteration on the performance of the participants. It means that two different groups were to take two various formats of the same test in four testing occasions. Hence, two groups were determined to take two versions of the same test in four occasions. First, testing group one took

PBT form of Vocabulary in Use Pre-intermediate and Intermediate Level Test and testing group two took the CBT version of the test, simultaneously. The conversed testing mode order was administered to two testing groups after four weeks interval to mitigate the practical potential, fatigue effects and testing effects. After the interval time, when testing group two (henceforth TG2 in tables) was taking the PBT form of the test in one of the classes of Faculty of Management and Humanities of CMU, testing group one (henceforth TG1 in tables) was taking CBT form of the same test in computer laboratory of Information Technology Center of CMU, simultaneously.

To investigate if any change occurred in test takers' preference toward taking paper-based or onscreen test, test takers were asked to answer a simple question mentioned at the bottom of their exam paper and screen, i.e. *would you prefer taking test on which mode? 1. on paper 2. no difference 3. on computer* to examine the relationship between testing mode preference and performance. This question studied the possible change in test takers' preference after taking CBT version of the exam. After implementing each CBT version of the test, participants were asked to fill out the CAS questionnaire. Besides, to confirm questionnaires data, 40 participants who filled out the questionnaire were randomly selected from among the volunteers for interview.

In fact, test takers' perceptions of PBT and CBT were examined with a post-test survey. This post-test survey was conducted in the form of forty one-to-one audio recorded interviews with the volunteers. The researcher decided to carry out interview to give the opportunity to the respondents to elaborate on their responses given to the simple testing mode preference questionnaire appeared at the bottom of exam paper and screen. In interviews, the respondents had the opportunity to elaborate on the reasons why they preferred a particular administration mode of testing and state their opinions about some particular test features they liked in both versions of the test.

One week after termination of CBT and PBT exams, 20 test takers from testing group one were randomly selected and then they were called and invited to be present in the Information Technology Center of CMU for the day after the call-day at the time determined previously. The approximate one week interval between the last test and the interview helped respondents feel free to state their opinions and reasons for their testing mode preference easily and answer the interview questions in parallel to the given responses to the questionnaire. One day after conducting interview with testing group one, the 20 randomly selected participants of testing group two were called and invited to be present in the Information Technology Center of CMU for the day after the call-day. 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, impressions of the CBT version of the test, their feelings about two testing modes, and their reasons to prefer a test modality. Some of the participants who changed their preference were also asked about their reasons to change their preferences after taking CBT.

4. Data analysis and results

For the first analysis, the one-way analysis of variance (ANOVA) was used to determine whether there was any statistically significant difference between the means of two independent (unrelated) groups. It was also used to

compare the means of two sets of scores of each group (related group) obtained in two different testing sessions. But, first, descriptive statistics are used to gain a better view of the data. The one-way ANOVA descriptive statistics output is displayed in Table 1.

Table 1: Descriptive statistics of two groups' mean scores in four testing sessions

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
TG1 PBT	57	**44.84**	**12.21**	1.61753	41.6018	48.0824	14.00	66.00
TG1 CBT	57	**47.08**	**19.59**	2.59558	41.8882	52.2873	14.00	98.00
TG2 PBT	57	**47.17**	**18.51**	2.43141	42.3036	52.0412	14.00	98.00
TG2 CBT	57	**40.00**	**14.70**	1.94794	36.0978	43.9022	20.00	98.00
Total	228	44.78	16.68	1.10238	42.6139	46.9582	14.00	98.00

According to the results, testing group one mean score on CBT (M = 47.08, SD = 19.59) was higher than that group's mean score on the PBT (M = 44.84, SD = 12.21). Testing group two mean score on PBT (M = 47.17, SD = 18.51) was higher than that group's mean score on the CBT (M = 40, SD = 14.70). Additionally, of the two CBT sessions of the test taken by testing groups one and two, the highest mean score was found for testing group one on CBT; with a relatively higher mean score by 7 points (Table 1).

On the other hand, the standard deviation in testing group one CBT was higher than in PBT. It means that the dispersion of scores from mean score in CBT was higher than in PBT; consequently, it was concluded that Standard Error of Measurement (SEM) in testing group one PBT was lower than in that group's CBT.

According to the output of One-Way ANOVA analysis in which the mean difference was significant at the 0.05 level, the significance level was .072 (i.e., p=.072) which was greater than 0.05. Therefore, it was concluded that there was no statistically significant difference in the mean score of two testing groups in four testing sessions as a whole (F (3,228) = 2.363, p=.072) (Table 2).

Table 2: ANOVA results comparing testing sessions of two testing groups

	Sum of Squares	D.F.	Mean Square	F	Sig.
Between Groups	1938.099	3	646.033	2.363	**.072**
Within Groups	61512.416	225	273.389		
Total	63450.515	228			

In addition to the Tukey HSD post hoc test in which the variances are assumed equal, the Tamhane's T2 and Games-Howell post hoc tests applied for the variances that are not assumed equal were the preferred ones from

among some others to do a multi-comparison. It was seen that there was no statistically significant difference between testing groups taking PBT and CBT versions of the test as determined by one-way ANOVA (F(3,225) = 2.363, p = .072). According to the results of Tukey HSD post hoc test, the differences between PBT and CBT versions of testing group 1 (p=.887), CBT versions of two testing groups (p=.104), and PBT and CBT versions of testing group 2 (p=.095) were not statistically significant (Table 3). *Tamhane's T2* post hoc test also displayed that there was no statistically significant difference between mean score of PBT and CBT versions of testing group one, mean score of PBT and CBT versions of testing group two, and between mean score of CBT version of testing group one and CBT version of testing group two that were indicated by (p =.976), (p = .132), and (p = .173) (Table 3). Additionally, *Games-Howell* post hoc statistical test for examining variables of the groups whose variances were not assumed to be equal revealed that there was no statistically significant difference between PBT and CBT versions of testing group one, PBT and CBT versions of testing group two, and between mean score of CBT version of testing group one and CBT version of testing group two that were indicated by (p = .883), (p = .104), and (p = .134), respectively (Table 3).

Table 3: Post Hoc tests results of four testing sessions with each other

		(I)Testing Sessions	(J)Testing Sessions	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Equal Variances Assumed	Tukey	Testing_Group1_PBT	Testing Group1_CBT	-2.24561	3.09719	**.887**	-10.2622	5.7709
		Testing_Group1_CBT	Testing_Group2_CBT	7.08772	3.09719	**.104**	-.9288	15.1043
		Testing_Group2_PBT	Testing_Group2_CBT	7.17241	3.08381	**.095**	-.8095	15.1543
Not Equal Variances Assumed	Tamhane	Testing_Group1_PBT	Testing_Group1_CBT	-2.24561	3.05834	**.976**	-10.4655	5.9743
		Testing_Group1_CBT	Testing_Group2_CBT	7.08772	3.24523	**.173**	-1.6163	15.7918
		Testing_Group2_PBT	Testing_Group2_CBT	7.17241	3.11548	**.132**	-1.1771	15.5219
	Games-Howell	Testing_Group1_PBT	Testing_Group1_CBT	-2.24561	3.05834	**.883**	-10.2453	5.7541
		Testing_Group1_CBT	Testing_Group2_CBT	7.08772	3.24523	**.134**	-1.3859	15.5613
		Testing_Group2_PBT	Testing_Group2_CBT	7.17241	3.11548	**.104**	-.9571	15.3019

Furthermore, according to the results of Multiple Comparisons output resulted from Tukey HSD in which equal variances are assumed and even *Tammany's T2* and *Games Howell* post hoc tests in which equal variances are not assumed (Table 3), groups did not differ from each other in their CBT performance. It means that testing order session might not be considered as a factor affecting testing performance.

To explore the interaction between test takers' prior computer attitudes and testing performance on CBT, Pearson Correlation was chosen to investigate the degree to which computer attitude influence computerized test performance. The results for testing group one (r (112) =.135, P >.05) indicated that there was no significant

relationship between the two variables (Table 4) and computer attitude was not statistically significant predictor of CBT performance. The results of the present analysis established that while computer attitude variable had a weak positive correlation with the testing performance, it was not dominant factor in determining computer test scores. According to the results, for the testing group one, the answers of participants to the first factor and their testing performance was not strongly correlated, .135 (112) =.405, P >.05.

Table 4: Pearson Correlation of computer attitude construct with CBT scores of testing group one and two

Testing Group 1 CBT	**Pearson Correlations** Attitude Construct	
	Pearson Correlation	**.135**
	Sig. (2-tailed)	**.405**
	N	114
Testing Group 2 CBT	**Pearson Correlations** Attitude Construct	
	Pearson Correlation	**.060**
	Sig. (2-tailed)	**.715**
	N	114

The Pearson Correlation that was run to determine the relationship between computer attitude external moderator variable and CBT score of testing group two indicated that there was no statistically significant correlation between this construct and CBT performance of testing group two (r (112) = .060, p = .715) (Table 4). According to the results, for testing group two, the answers of participants to the first factor and their testing performance was not strongly correlated, .060 (112) =.715, P >.05. Responses to the simple question appeared at the bottom of PBT version of testing group one were correlated with participants' mean score on computerized test to see if there was any significant correlation between their prior testing mode preference and testing performance on CBT. Additionally, using descriptive statistics, we also performed multiple comparisons between three preference groups of each testing group to examine the relationship between the prior testing mode preferences and performance on computerized tests. A Pearson's product-moment correlation was run to assess the relationship between pre and post-CBT mode preference and CBT performance of all the test takers of testing group one. There was a weak positive correlation between both pre and post-CBT mode preference and CBT performance of testing group one, r (114) = .015, p < .817 and r (114) = .92, p < .380, respectively, that were not statistically significant (Table 5).

Furthermore, the Pearson's product-moment correlation to assess the relationship of post-CBT and post-PBT mode preference with CBT performance of all the test takers of testing group two revealed that there was a weak positive correlation between both post-CBT and post-PBT mode preference and CBT performance of testing group two, r (114) = .083, p < .454 and r (114) = .210, p < .580, respectively, that were not statistically significant (Table 5).

Table 5: Pearson Correlation of pre-CBT and post-CBT mode preference with CBT scores of testing group one and post-CBT and post-PBT mode preference with CBT scores of testing group two

Pearson Correlations		Pre-CBT Mode Preference	Post-CBT Mode Preference	Post-CBT Mode Preference	Post-PBT Mode Preference
CBT 1	Pearson Correlation	**.015**	**.92**		
	Sig. (2-tailed)	**.817**	**.380**		
	N	114	114		
CBT 2	Pearson Correlation			**.083**	**.210**
	Sig. (2-tailed)			**.454**	**.580**
	N			114	114

To examine the relationship between testing mode preference and testing performance, the following descriptive data was also used.

Table 6: PBT performance of different preference groups of testing group one on PBT

Pre-CBT Mode Preference	N	**PBT 1 Mean Score**	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
On Paper	75	**40.57**	8.34	1.57	37.3368	43.8061	28.00	52.00
No Difference	20	**47**	1.06	.377	46.1063	47.8937	46.00	48.00
On Computer	19	**59.85**	8.73	.976	57.9070	61.7930	32.00	66.00
Total	114	44.20	9.98	1.57	41.0074	47.3926	28.00	64.00

As it was shown in the Table 6, the PBT mean score of On Computer preference group (PBT1 / M = 59.85, (SD = 8.73)) was higher than the other two preference groups. It means that the test takers of testing group 1 who preferred CBT over PBT did better than those who preferred PBT (PBT1 / M = 40.57, (SD = 8.34)) on PBT version of the test. On the other hand, those who expressed their preference as taking the PBT version of the test in PBT testing session had better performance on CBT testing session (CBT1 / M = 41.42, (SD = 15.95)). But, the test takers who preferred taking the test on CBT version, did not better on their preferred testing mode (Table 7).

Table 7: PBT performance of different preference groups of testing group one on CBT

Pre-CBT Mode Preference	N	**CBT 1 Mean Score**	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
On Paper	28	**41.42**	15.95	3.01	35.2429	47.6142	14.00	66.00
No Difference	8	**56**	4.27	1.51	52.4250	59.5750	52.00	60.00
On Computer	4	**59.11**	9.20	1.11	56.8884	61.3469	32.00	66.00
Total	40	**46.6**	15.74	2.48	41.5652	51.6348	14.00	66.00

To compare the results of different testing mode preference groups of testing group one on PBT and CBT sessions (between groups), as the Table 6 revealed, those participants who preferred taking PBT version of the test (PBT1 / M = 40.57, (SD = 8.34)) outperformed on their CBT exam (CBT1 / M = 41.42, (SD = 15.95)) (Table 7). Accordingly, those who preferred taking the test on CBT (CBT1 / M = 59.11, (SD =9.20)) (Table 6), before implementing CBT version of the test, had better performance on their PBT exam (PBT1 / M = 59.85, (SD = 8.73)). And those who didn't mind taking the test on either mode (PBT1 / M = 47, (SD = 1.06)), did better on CBT (CBT1 / M = 56, (SD = 4.27)) (Table 7). However, the overall results 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 [35].

According to the results, after implementing PBT version of the test and before the second testing administration session, more than 65% of the test takers of testing group one preferred to take the test on paper, and 17.54% of the test takers didn't mind taking the test in either mode and just 16.66% opted for computers as their preferred mode of testing (Table 8).

To see if any change has happened to the mode preference of test takers of group one, their answers to the second simple questionnaire were examined. As it was shown in the Table 8, only 13.15% of the test takers still preferred PBT version of the test, while just 11.40% didn't mind taking the test on either mode. The greater percentage 75.43% was the test takers who opted for computer as their preferred mode of testing. According to the results of Table 6, we concluded that the number of participants who preferred PBT and who didn't mind

taking the test in either mode in PBT1 testing session have changed in favor of the test takers who chose On Computer as their preferred testing mode preference after exposure to CBT version of the test. It means that exposure to CBT version of the test developed positive attitudes towards it.

Table 8: Frequency Table of responses to the Pre-CBT Post-CBT testing mode preference of testing group one

Pre-CBT 1 Testing Mode Preference				
		Frequency	Percent	Valid Percent
Valid	On Paper	75	65.78	65.78
	No difference	20	17.54	17.54
	On Computer	19	16.66	16.66
	Total	114	100.0	100.0
Post-CBT 1 Testing Mode Preference				
		Frequency	Percent	Valid Percent
Valid	On Paper	15	13.15	13.15
	No Difference	13	11.40	11.40
	On Computer	86	75.43	75.43
	Total	114	100.0	100.0

According to the responses of 20 post-CBT simple questionnaires of testing group one and the responses of 20 post-PBT simple questionnaire of testing group two who were invited for interview, 82.5% (33 people) preferred computerized test and 17.5% (7 people) showed preference for paper-based test.

In the interview, 33 people who advocated the CBT and 7 ones who preferred PBT were asked some questions to rationalize their testing mode preference and explain their reasons to find out the rationales behind each preference to support the findings of quantitative analysis. The recorded statements of respondents were transcribed and content analyzed by two TEFL experts of CMU and then the transcribed data were classified thematically under two categorizations including 15 Preferred Features and 3 Not-Preferred Features of CBT testing mode and 7 Preferred Features and 3 Not-Preferred Features for PBT testing mode (Appendix A).

Based on the results, most of the participants showed high CBT preference as well as more advantages for CBT over PBT to rationalize why they preferred this mode of testing. It was concluded that the participants' answers to the interview questions were in line with their responses to the simple questionnaire on their preferred testing mode. Among those who responded the interview questions, 100% of the participants who favored CBT version of the test mentioned "Easy to read items", "Easy to choose answers", Easy to change answers", and "Immediate scoring reports" as their reasons to prefer CBT mode of administration.

More than 70% enjoyed CBT testing environment and 65.56% liked CBT because it was less fatiguing. Furthermore, 87.62% of CBT advocators of CBT believed that CBT was more comfortable, and 69.15% liked it

because it was faster testing mode with fewer recognition errors. 52% liked CBT due to its less time needed to revise the item question and shorter time to response the question. More than 78.24%, 60.85%, and 57.45% of the CBT advocators had positive attitudes towards the CBT features including "Enhanced security", Faster decision making as a result of immediate scoring and reporting", and "less time and effort" to take this format of the test, respectively.

Moreover, the reason of 30.75% of test takers to advocate CBT was due to its accuracy while more than 90% of the CBT advocators favored CBT because no human error could have impact on their test results. Additionally, more than 93% of the participants justified that since conventional tests' format and the way that the papers are distributed to the students are boring, they didn't like that format of the test.

Despite the high percentage of CBT preference that were reported by the respondents of the interview questions, some of the participants still preferred the conventional format of the test. Among the advocators of PBT, 100% selected "Easy to navigate", "More familiarity with testing format and conditions", "Being accustomed to circle the questions and answers for later review", and "No need to extra task demand" as the advantages of PBT and their reasons to advocate this format of the test. They also declared that reviewing the answers was time consuming in CBT (85.71%) because just one question was displayed in the screen and it was time consuming to go back to question 1 if they were on question 35, for example. "Requiring technical knowledge" (57.14%) and "Concern of system breaking down" (71.42%) were among the other disadvantages that the participants selected as the Not-Preferred Features of CBT version of the test.

5. Conclusion

The purpose of this study was to compare two sets of scores received from two modes of testing administration i.e. PBT & CBT to determine whether computerized testing affected student's achievement. In fact, mode effect on testing performance of test takers is investigated. Mode effect is defined as a discrepancy that is recognized between the PBT and CBT testing modes' performance. Clariana and Wallace define mode effect as "the empirical evidence that identical paper-based and computer-based tests will not obtain the same results" (p. 593) [21]. For the first research question, the means of two sets of scores of two testing groups (related group) and (unrelated groups) obtained in four different testing sessions was explored. Based on the findings, it was concluded that there was no statistically significant difference in the mean scores of two testing groups in four testing sessions as a whole (p=.072). The findings of the research question one were compatible with the results of [36,37,38,39] who claim that assessments are comparable across modes. The findings were also in contrast to the other researchers [40,29] who disagree with the comparability of scores obtained from two testing modes. By considering comparability studies in Iranian educational contexts, unlike [24] and [41], the findings of this study are in line with the findings of [42] that supports the equivalency between test scores of PBT and CBT. To answer the research question 1.1, Tukey HSD test showed that the difference between two CBT versions of two testing groups was not statistically significant (*Sig* = .104, p>0.05). Besides, Tamhane (*Sig* = .173, p>0.05) and Games-Howell (*Sig* = .134, p>0.05) post hoc tests didn't show any statistically significant difference between CBT performance of testing group one that was implemented in the second testing session and CBT performance of testing group two that was implemented in the first testing session. Based on these findings, we

concluded that if the order of test was reversed, the students would still receive the same scores and alteration of test order was not considered a factor influencing CBT performance. The second research question that tried to find the relationship of computer attitude with CBT performance was analyzed by Pearson Correlation. For CBT performance of testing group one, the index of correlation was P=.135 for attitudes towards computer. On the other hand, for CBT performance of testing group two, the index of correlation was P=.060 for attitudes towards computer. The findings went with other similar studies such as [43]. Like this study, he found no statistically significant correlation or interactive positive effect between computer attitude and CBT performance. For analyzing research question three which focused on testing mode preference, the results revealed that there was no statistically significant correlation between testing mode preference of test takers before and after CBT version of the test and their CBT testing performance in two testing groups. The findings supported the previous research done by [44] and [10] in which there was a high preference for CBT, but test takers' preference had negative correlation with their performance on CBT. Similar to the present study, no significant difference was found between test takers' scores on two versions of test which indicated no correlation between test mode preference and test performance [45]. Therefore, we reached to the conclusion that although exposure to the CBT may change prior testing mode preference and may lead to positive attitudes towards this kind of test version, the prior testing mode preference as an external moderator factor does not have influence on the CBT testing performance of the participants. These findings indicated that there was no necessarily positive interaction between testing mode preference and testing performance. The findings of the present study were in consistent with the result of [39] study that found out test takers with positive attitudes towards the use of computer did not perform better on CBT. According to the qualitative research data, most of the participants showed high CBT preference as well as more advantages for CBT over PBT to rationalize why they preferred this mode of testing.

The results indicated higher preference rate for CBT but better performance on PBT. It can be concluded that the participants' answers to the interview questions were in line with their responses to the simple questionnaire on their preferred testing. These finding support previous studies in the literature. Boo found that their participants showed more preference towards CBT [46]. His respondents claimed that PBT was more comfortable and less fatiguing than CBT while CBT was easier for them to record and change their answers. The finding of the present study and Boo's study are mostly in line with [43] whose participants showed more preference on computerized tests but performed better on paper-based tests. The information of testing mode preference and attitudes towards CBT and its features were supported by focus group interview conducted in this research. Some inevitable limitations and difficulties would pose themselves upon the research process while it is being carried out. Accordingly, there were several limitations to the current study. First, although some variables such as testing mode effects, testing mode order effects, prior attitudes towards computer, and testing mode preference are considered in this study, many other related external moderator variables such as computer familiarity, computer anxiety, ethnicity, intelligence, affective and motivational factors, test anxiety, test effects, testing comfort levels, differences in testing conditions, cognitive processing, characteristics of computers being used, screen size and resolution, font characteristics, line length, number of lines, interline spacing, white space, scrolling, item review and item presentation that may influence the measured performance of the test takers were not taken into account here due to the limitations of every experimental study. It is

recommended for further research on these issues. Second, since the resulted findings are specific to the test of vocabulary knowledge of ESP students of CMU administered at the end of the second educational semester in Chabahar Maritime University of Iran in 2016-17 educational year, then it cannot be generalized to the other language skills and sub-skills and also other CBT testing programs. Then, the results of the study are valid only in relation to the English vocabulary knowledge of ESP students. Furthermore, the results of this study should not be generalized to the contexts in which the participants are more heterogeneous in knowledge and ability. Then, further replications of the study with more participants who are less homogeneous would be desirable thereafter. The next limitation is that since CBT testing programs depend heavily on linear item selection algorithm, the results that were obtained and presented in the present study are specific to the CBT testing programs which use this kind of item selection algorithm. In fact the results are not applicable to the other computerized testing programs such as CAT with adaptive kind of algorithm to select and present test items. Then the number of university students who participated in the study was not large. Only undergraduate students of Chabahar Maritime University were involved in this study. Then, it might not be possible to generalize the results of the study to the entire population of learners of English as Specific Purposes due to the characteristics of the sample groups and issues related to diversities in many fields such as sex, age, race or etc. And finally' some other applicable expected limitations for this study including Hawthorn effect and research resistance might occur too. The Hawthorn effect influences the test result by students' responding under the influence of being a participant in a study. Research resistance may have come into the effect if students resist against the research. For example they may participate in the study but actually they are resented by losing their recess or lunch hour because of taking the test. These psychological aspects may influence test results. Actually, the limitations of computer analyses of human language did not allow us to address directly the more important assessment of communicative competence. Additionally, in conjunction with the linear model of computerized multiple-choice vocabulary knowledge test, the proposed study was confined to the linear scoring approaches to examine the score equivalency of CBT and PBT. More sophisticated approaches are suggested to be used in the future studies. Although some variables such as computer anxiety, prior computer familiarity, and gender as well as many other related variables such as ethnicity, intelligence, affective and motivational factors, differences in testing conditions, cognitive processing, characteristics of computers being used, screen size and resolution, font characteristics, line length, number of lines (and some others as mentioned in the limitation section) that may influence the measured performance of the participants are recommended for further research. Another suggestion is to test other language skills such as reading skill in a more comprehensive study in order to widen the insights to the language testing in comparability studies.

Acknowledgements

I am most grateful to the members of Language Department of CMU for their time, expertise and guidance integral to the completion of this dissertation. I am especially grateful to ***Dr. Hooshang Khoshsima*** whose attention to the details, patience and knowledge while guiding me through this process, support, encouragement, expertise and kindness helped me throughout the present research project. I would like to give a heartfelt thank you to ***Vahide***, my wonderful wife, for sticking by me through all the time I was engaged in the research completion.

References

[1] Challoner, J. (2009). 1001 Inventions that changed the world (Cassell Illustrated: 2009).

[2] 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.

[3] American Psychological Association (APA). (1986). Guidelines for computer-based tests and interpretations. Washington, DC: Author.

[4] 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.

[5] Fleming, S., Hiple, D., (2004). Distance Education to Distributed Learning: Multiple Formats and Technologies in Language Instruction. CALICO Journal, 22 (1), 63-82.

[6] Bennett, R.E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. The Journal of Technology, Learning and Assessment, 1(1), 1-24.

[7] 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).

[8] Peat, M. & Franklin, S. (2002). Supporting student learning: the use of computer-based formative assessment modules. British Journal of Education Technology, Vol. 33, No. 5. https://doi.org/10.1111/1467-8535.00288.

[9] Zhang, L., & Lau, C. A., (2006). A comparison study of testing mode using multiple-choice and constructed-response items – Lessons learned from a pilot study. Paper presented at the Annual Meeting of the American Educational Association, San Francisco, CA.

[10] 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(2017). http://dx.doi.org/10.5539/elt.v10n2p23.

[11] 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.

[12] 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.

[13] Scherer, R., & Siddiq, F. (2015). The big-fish-little-pond-effect revisited: do different types of assessments matter? Computers & Education, 80, 198e210. http://dx.doi.org/10.1016/j.compedu.2014.09.003.

[14] 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.

[15] Nikou, S. A., & Economides, A. A. (2013). Student achievement in paper, computer/ web and mobile based assessment. In Proceedings of the 6th Balkan Conference on Informatics (BCI), Greece.

[16] 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].

[17] Chua, Y. P., & Don, Z. M. (2013). Effects of computer-based educational achievement test on test performance and test takers' motivation. Computers in Human Behavior, 29(5), 1889e1895. http://dx.doi.org/10.1016/j.chb.2013.03.008.

[18] Norazah Mohd Nordin, N. M., Arshad, S. R., Razak, N. A., & Jusoff, K. (2010). The Validation and Development of Electronic Language Test. Studies in Literature and Language, 1, 1-7.

[19] Mazzeo, J., Druesne, B., Raffield, P. C., Checketts, K. T., & Muelstein, A. (1991). Comparability of computer and paper-and-pencil scores for two CLEP general examinations. College Board Report No. 91-5. New York. (ERIC Document Reproduction Service No. ED344902).

[20] Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. Education Policy Analysis Archives, 7, 20. https://doi.org/10.14507/epaa.v7n20.1999.

[21] 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.

[22] DeAngelis, S. (2000). Equivalency of computer-based and paper-and-pencil testing. Journal of Allied Health, 29(3), 161–164.

[23] 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.

[24] Hosseini, M., Zainol Abidin, M. J., Baghdarnia, M., (2014). Comparability of Test Results of Computer-Based Tests (CBT) and Paper and Pencil Tests (PBT) among English Language Learners in Iran. International Conference on Current Trends in ELT, 659-667.

[25] 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.

[26] Paek, P. (2005). Recent trends in comparability studies (PEM Research Report 05-05). Available from http://www.pearsonedmeasurement.com/downloads/research/RR_05_05.pdf.

[27] 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. tps://doi.org/10.1177/0013164406288166.

[28] 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. https://doi.org/10.1177/001316448604600224.

[29] Fulcher, G. (1999). Computerizing an English language placement test. ELT Journal, 53(4), 289-299. https://doi.org/10.1093/elt/53.4.289.

[30] Leeson, H. (2006). The Mode Effect: A Literature Review of Human and Technological Issues in Computerized Testing. International Journal of Testing, 6(1), 1-24. https://doi.org/10.1207/s15327574ijt0601_1.

[31] Loyd, B. H., & Gressard, C. (1984). Reliability and factorial validity of computer attitude scale. Educational and Psychological Measurement, 44(2), 501-505. https://doi.org/10.1177/0013164484442033.

[32] 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. https://doi.org/10.1177/0013164485454021.

[33] Berberoglu, G. & Calikoglu, G. (1992). The construction of a Turkish computer attitude scale. Studies in Educational Evaluation, 24 (2), 841-845.

[34] Christensen, R., & Knezek, G. (1996). Constructing the Teachers' Attitudes toward Computers (TAC) questionnaire. Paper presented to the Southwest Educational Research Association Annual Conference, New Orleans, Louisiana, January, 1996. (ERIC Document Reproduction Service No. ED398244.

[35] Al-Amri, S. (2007). Computer-based vs. Paper-based Testing: Does the test administration mode matter. Proceedings of the BAAL Conference, 2007.

[36] Al-Amri, S. (2008). Computer-based testing vs. paper-based testing: A comprehensive approach to examining the comparability of testing modes. Essex Graduate Student Papers in Language and

Linguistics, 10, 22–44. Retrieved January 28, 2012 from http://www.essex.ac.uk/linguistics/publications/egspll/volume_10/pdf/EGSPLL10_2244SAA_web.pdf.

[37] Green, T. & Maycock, L. (2004). Computer-Based IELTS. Research Notes, Issue 8, pp. 3-6.

[38] Hashemi Toroujeni, S.M. "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 Art in Teaching. Chabahar Marine and Maritime University (Iran) (2016).

[39] Khoshsima, H. & Hashemi Toroujeni, S.M. (2017). 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 (2017). http://dx.doi.org/10.5281/zenodo.268576.

[40] Coniam, D. (2006). Evaluating computer-based and paper-based versions of an English language listening test. ReCALL, 18, 193-211. https://doi.org/10.1017/S0958344006000425.

[41] Salimi, H., Rashidy, A., Salimi, A. H., & Amini Farsani, M. (2011). Digitized and non-Digitized Language Assessment: A Comparative Study of Iranian EFL Language Learners. International Conference on Languages, Literature and Linguistics, (IPEDR), vol.26. IACSIT Press, Singapore.

[42] Mojarrad, H., Hemmati, F., Jafari Gohar, M., and 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 Linguistic World, 4 (4), 418 428.

[43] Al-Amri, S. (2009). Computer based testing vs. paper based testing: Establishing the comparability of reading tests through the revolution of a new comparability model in a Saudi EFL context. Thesis submitted for the degree of Doctor of Philosophy in Linguistics. University of Essex (UK).

[44] 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. https://doi.org/10.1177/016264341102600102.

[45] 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.

[46] Boo, J. (1997). Computerized versus paper-and-pencil assessment of educational development: Score comparability and examinee preferences. Unpublished PhD dissertation, University of Iowa, USA.

Appendix A:

Table 9

	Themes of features			
	Preferred Features		Not Preferred Features of CBT	
Computer-Based Testing	Easy to read items	1	Reviewing the answers is time-consuming	1
	Easy to choose answers	2	It needs technical knowledge and IT skills	2
	Easy to change answers	3	system breaking down concern	3
	More enjoyable	4		
	Less fatiguing	5		
	More comfortable and flexible environment	6		
	Faster test taking	7		
	Fewer response entry and recognition errors	8		
	Faster and more controlled test revision process with shorter response time	9		
	Instant score report	10		
	Enhanced security	11		
	Faster decision-making as the result of immediate scoring and reporting	12		
	Less time and effort	13		
	Decrease in testing error	14		
	No human error in automated scoring	15		
Paper-Based Testing	Preferred Features		Not prefered Features of PBT	
	Easy to navigate	1	Boring test format	1
	Easy to review answers	2	Distribution of papers	2
	More familiarity with testing format and conditions	3	human error scoring	3
	Less risk of technology issues	4		
	Being accustomed to taking notes	5		
	Being accustomed to circle the options for later review	6		
	No need to extra task demand	7		