International Journal of Language and Linguistics
2017; 5(3): 78-87
http://www.sciencepublishinggroup.com/j/ijll
doi: 10.11648/j.ijll.20170503.13
ISSN: 2330-0205 (Print); ISSN: 2330-0221 (Online)

Replacing Paper-Based Testing with an Alternative for the Assessment of Iranian Undergraduate Students: Administration Mode Effect on Testing Performance

Monirosadat Hosseini[1], Seyyed Morteza Hashemi Toroujeni[2]

[1]Department of English Language, Faculty of Humanities, Farahan Payame Noor University, Farahan, Iran

[2]English Language Department, Faculty of Management and Humanities, Chabahar Marine and Maritime University, Chabahar, Iran

Email address:

Ho.mahmonir@yahoo.com (M. Hosseini), Hashemi.seyyedmorteza@gmail.com (S. M. H. Toroujeni)

To cite this article:

Monirosadat Hosseini, Seyyed Morteza Hashemi Toroujeni. Replacing Paper-Based Testing with an Alternative for the Assessment of Iranian Undergraduate Students: Administration Mode Effect on Testing Performance. *International Journal of Language and Linguistics*. Vol. 5, No. 3, 2017, pp. 78-87. doi: 10.11648/j.ijll.20170503.13

Received: April 15, 2017; **Accepted**: April 26, 2017; **Published**: May 24, 2017

Abstract: There have been studies on comparability of test results in Computer-Based Testing (Henceforth CBT) and Paper-Based Testing (Henceforth PBT) considering key factors associated with test results in different countries with different languages and technological backgrounds. The main purpose of the current study was to discover the equivalency of test scores on PBT and CBT in the English achievement test in Payame Noor University (PNU) among undergraduate students. It also intended to investigate if there was any relationship between computer attitude and testing performance on CBT. Based upon the quantitative and qualitative data, some major findings were revealed. Firstly, there was statistically significant difference between two sets of mean scores. Furthermore, based on descriptive results, in comparing the results of computerized and paper-based tests, students showed better performance on PBT than CBT. The results of this study support the necessity of doing comparability studies in higher educational contexts before substituting CBT for PBT or including it in the system. Then, computer attitude had not any interaction with testing performance on CBT among Iranian undergraduate students in PNU. Finally, the results of interview supported the quantitative findings, i.e. participants mostly showed high preference for computerized test and liked CBT more than PBT but due to some justifications and habit of taking tests traditionally, they performed better on PBT.

Keywords: Computer-Based Testing, Paper-Based Testing, Testing Administration Mode, Computer Attitude

1. Introduction

One of the most appropriate ways of measuring students' learning in educational setting is assessment [2, 3]. Portfolio assessment, performance assessment, self-assessment, and peer assessments are among the examples of different types of assessment [4]. In recent years, information and communication technology has been employed in assessment and examination to mechanize the testing process. Computer based testing (CBT) provides a variety of innovations in testing and can be used in different contexts, one of the important areas is language testing [5]. The history of computerized testing began in the early 1970s [2, 6, 7, 8, 9, 10]. With the appearance of new technologies, computerized testing has begun to be implemented in large scale testing [11, 12]. Examples include state drivers' license exams, military training exams, job application exams, entrance exams in postsecondary education, and certification exams by professional groups such as TOEFL or IELTS [13, 14]. The limited accessibility to computer and high cost limited the implementation of computerized language testing in past years [15, 16, 17]; however, recent developments in communication technologies have created alternative test methods through computers and internet all around the world [1].

The necessity of using technological devices both in learning and testing in educational settings have been rapidly increased since widespread accessibility to computers and broad developments in information and communication technology [2, 7]. In this regards, computer based testing is

going to be applied all around the world in academic contexts [16, 1, 3, 18]. Developments in language testing studies during the past years provided progress evidence in making use of technological devices in education and in language testing, mostly in test development, administering, storage, scoring, reporting and processing given data [19]. Meanwhile, Factors such as increased accessibility to personal computers, widespread use of computer in language learning, increased computer familiarity and positive attitude towards the use of computer in educational settings motivated researchers to conduct studies considering these influencing factors in comparability studies [20, 3, 12, 21, 22, 23, 24].

Given the integral role computers play in our lives, the number of computer-delivered tests is increasing in language testing due to the perceived advantages of computer-delivered tests [17]. Such developments in computer technologies have influenced many areas including educational settings such as online learning and testing [25, 26, 27]. In language learning, the use of computers and electronic devices has become popular; especially in assessing the language proficiency of English learners, the most precise and available way is through computers [2, 1, 17, 28]. However, the limited accessibility to computer and high cost of using computer in high stake tests had limited the implementation of computerized language testing [1].

This study aimed at examining the score comparability of multiple-choice General English achievement tests in two different testing modes, paper-based tests (PBT) and computer-based tests (CBT) taken by Iranian students of Payame Noor University (PNU). It also aimed at finding out the relationship of test takers' attitude towards the use of computer. Therefore, the investigators tried to introduce a comprehensive theoretical model of examining score comparability and establish an outline for implementing English language computerized tests in Iran considering effective factors influencing test performance on CBT such as attitude towards the use of computer.

2. Literature Review

Computer-based test is a critical part of any web-based tests and the issue of comparability is considered very important for test developers and curriculum designers when deciding to substitute CBT for PBT or include it in their programs.

Although computer-based testing has advantages over paper-based testing [29, 1, 3, 30], equivalency of two test modes should be ensured first [1, 17, 28, 31]. Due to the importance of this issue, the American Psychological Association (APA) assigned guidelines for CBT and its interpretations to retain the equivalency with PBT. Choi, Kim, and Boo defined equivalency as an investigation into the comparability of test modes or test tasks represented in different testing conditions [32]. Neuman and Baydoun recommended equivalency of tests as: "the extent to which different formats of the same test measure the underlying trait determines whether they can replace each other" [33].

Reviewing related literature on the comparability studies on CBT and PBT shows different results and opposite findings. In some of these studies, the test scores of two tests were similar [15, 34, 35, 3, 30, 12, 36]. Some other, in contrast, found different results with the priority of CBT over PBT and vice versa. For example, some studies that showed higher score on CBT include [1, 37]. Contradictory findings reported lower performance on CBT than PBT [38, 39, 40, 41].

Perhaps one of the most important reasons in the differences in test results in relation to test mode effects of PBT and CBT is the difference in flexibility of test modes. Probably, it is because some CBTs do not provide the same level of flexibility as PBTs provide or vice versa. For example, some computer interfaces do not allow the student to skip, review, or change answers [1]. Mason and his colleagues also found evidence that shows the influence of different levels of flexibility on test results [42]. There have been numerous works on the effect of changing answers on PBT results, and the results demonstrate that changing answers on multiple-choice tests in CBT slightly increases scores [43, 44, 45].

However, some of these researchers attributed the differences to the similarity of the two test delivery modes. If computer-based tests closely are similar to paper-and-pencil format ones, the results could be similar as well. Evidence has accumulated to the point where it appears that in many traditional multiple-choice test settings, the computer may be used as a medium to administer tests without any significant effect on student performance‖ [17]. Any differences on multiple-choice tests, regarding the constructed response assessments, are related to individual basis. While most students prefer using the computer to paper format, their scores often vary depending on the mode of the test presentation [38, 46]. National Assessment of Education Progress (NAEP) in the Math online study suggests that performance on computer-based test items depends on the level of students' familiarity with using a computer. Students who are more familiar with the computer and more skillful in typing are more likely to perform better on the computer-based test. This finding suggests that computer familiarity may distort the measurement of mathematics achievement when tests are administered online to students who lack basic technology skills [47]. This conclusion motivated researchers to examine the correlation between familiarity and test result on CBT while comparing two test modes.

There are several investigations regarding the comparability of test scores of computer-based tests and paper-based test among students in many fields of studies [39, 48, 49, 17, 50]. Many focused on the differences between computerized tests and traditional paper and pencil tests without considering the effects of the learner adequately, if at all. Fletcher and Collins listed the advantages and disadvantages of CBT over PBT without addressing the effects of learners' characteristics on test performance [51]. On the other hand, there are few studies on the comparability of test modality of CBT and PBT in General English

Language test performance in the form of multiple-choice form [17, 28, 1].

Anakwe has done a study on the comparison between test scores of online testing and paper-based testing. The participants of his study were 75 undergraduate students each of them was given two tests, one in computerized format and one in paper and pencil format (within-subject grouping). The results of their data analysis showed no significant differences between test scores of participants on two test modes [15].

Similar to Anakwe [15], Al. Amri [29] study on the comparison between test scores of CBT and PBT among undergraduate students in Saudi Arabia showed no significant difference between mean scores but slightly better performance on PBT. Al-Amri also used within-subject group to decrease the effect of individual differences on two test performance. To test the research hypotheses, he used several statistics including, t-test for comparing test results, correlation analysis to find out the relationship between named variables and CBT scores, ANOVA for comparing test results of different groups, and content analysis for analyzing qualitative data. He, along with testing the difference between test scores of PBT and CBT, examined the effect of attitude towards the use of computer, familiarity with computer, prior and post preference for test mode, and test taking strategies on computerized test performance. He found no significant relationship between attitude and familiarity with computer and test performance on CBT. However, he could demonstrate the need of different test taking strategies in CBT than in PBT.

The main concern in using CBT in educational setting is the validity of tests. Messick described validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales supports the adequacy and appropriateness of inferences and actions based on test scores or other mode of assessment " [52]. Chapelle refers to validity as the test score interpretation and its use in a particular context [53]. Johnson and Green suggest that if CBT is going to be helpful in fulfilling the need of its users, it should match the level of validity and reliability of its counterpart mode, PBT [47]. This is why many researchers have conducted studies to examine comparability of PBT and CBT regarding the validity and reliability of CBT [29, 54]. It is still necessary to do more study regardless of advantages of CBT to examine the effect of converting tests from PBT to CBT on test performance of test takers [11]. Fulcher went further and suggested not only the equivalency of two test modes should be considered in comparability studies, but also other influencing factors such as prior computer experience, test takers' attitudes towards such systems, and liking, should be examined as well [55]. McNamara also suggests that the aim of validity of language testing is mostly to ensure the defensibility of fairness of interpretations based on test performance. He added that validity is a matter of degree, not all or none [56].

However, some studies have been conducted on the relationship of computer familiarity and attitudes of students with their test performance on CBT in their comparability studies of CBT and PBT [57, 29, 54, 1, 21, 58, 59, 60]. Leeson, in his study on the performance of students in computerized tests, identifies the factors that lead to difficulties in CBT performance. He identified the factors of the size and resolution of monitors, the way the questions and items presented through computer, writing characters and its fonts and length, the option of reviewing or changing the answers as those factors related to the 'technology used'. Whereas the user's gender, age, the ability to process information, familiarity with computer and ability to use it, the level of anxiety, and attitude are considered as user's' originated factors [61]. Some studies done on investigating the relationship between computer usage ability and achievement have emphasized that computer usage skill is an important predictor of user's success and students with poor computer ability show low achievement in CBTs [62, 41, 63, 64]. However, they stress that with the increase accessibility to and familiarity with computers, such problems may decrease. Thus, the present study attempts to confirm their claims by considering computer familiarity as a related factor in student's achievement computerized tests.

Kutluca and Gokalp have done a study on the computer usage and attitudes towards the use of computer of prospective preschool teachers. The result of this study, using "Computer attitude Scale" (CAS), showed that there was significant relationship between computer usage and attitude of using computers [65]. The result of the study done by Khoshsima and Hashemi Toroujeni on the attitudes of students and teachers towards the use of computer technology in Geography education, employing the CAS, revealed that there was no relationship between attitude of students towards using computer and developing curriculum [21].

Hashemi Toroujeni conducted a similar study on the students' attitude towards the use of computerized test among 80 undergraduate students in Iran with the aim of identifying the perceptions of students towards such systems. For research purpose, he examined the attitude of participants towards CBT and found out positive attitude towards the use of computer among the participants. In fact, he tried to find the relationship of computer attitude with CBT performance by running Pearson Correlation. Based on his findings, he found no statistically significant correlation or interactive positive effect between computer attitude and CBT performance [3].

Rezaee, Abidin, Issa, and Mustafa reviewed different theories related to factors influencing the success in accepting and applying technological tools in language teaching contexts. The aim of their study was finding out the relationship between teacher in services' attitudes towards the use of computer and cultural perception and computer competence. They found that having higher computer attitudes towards the use of computers lead to a higher perception of using computers. They also found that cultural perception was highly related to attitude towards the use of computers. According to TAM, cultural perceptions can be

identified as external factors that indirectly influence the behavioral intention and actual use. On the other hand, computer competence, as the other external factor influencing the actual use and directly influencing the attitude, was shown to have a positive effect on the participants' attitude towards the use of computers [66].

Powers and O'Neill examined the effect of computer attitudes (in their study referred to as testing mode preference) among other variables such as computer experience and computer anxiety which includes 'computer liking' (in their study referred to as computer attitudes). They employed questionnaire of computer attitudes (CAS), and a background questionnaire. Their findings showed non-significant relationship between these variables and test performance of participants. In contrast, the respondents showed more positive attitude towards CBT than PBT after their experience [67].

3. Methodology

3.1. Participants

In order to investigate the effect of testing administration mode on test takers' performance a total of 60 undergraduate students of Tehran Branch of PNU (Payame Noor University) in Iran were chosen. Of the 60 homogeneous participants, there were slightly more females (n=57%) than males (n=44%). The age of students ranged from 19 to 23 whose average age was 21.6 (M=60, SD=1.20). These students were of different majors but 12 of them were majoring in English. The 60 homogenous students were randomly selected and organized into one testing group to participate in two testing sessions in order to take both PBT and CBT versions of the same test.

3.2. Instruments

As one of the most important parts of any investigation is applying appropriate instrument to collect the required data, it is very essential for any researcher to choose as precise and appropriate tool as possible to collect adequate and related data [68]. Choosing inappropriate instrument would change the path of the study and lead to inappropriate and unrelated data [69]. This is why in this part, the researcher felt it necessary to elaborate the instruments and demonstrate their validity and reliability and report the pilot study for all instruments. Nelson Proficiency Test which was designed for a 30 (60%) pass mark was administered to determine the homogeneity and proficiency of the students.

The testing group participated in two testing sessions to take both formats of the same test derived from General English book on separate testing sessions with four weeks interval to mitigate the practical potential, fatigue 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. Comprehensive system of testing management of SAD was the program that was used to administer computerized testing. It is a test maker motor engine for developing any type of tests without limitations in the scope and administration. It is one of the most valid and widespread software developed in Iran, which had been tried out by other universities and organizations before. Another instrument for research data collection was a 30-item scale questionnaire entitled "Attitudes towards Computer-Based Testing" developed by the researchers themselves. To develop the researcher-made questionnaire, an item pool including 64 statements about the issues related to the CBT features and attitudes towards CBT was organized. Then it was presented to the experts from different areas such as computer and educational technology as well as educational testing and assessment. The 48 statement questionnaire was distributed to 120 undergraduate students to analyze its validity and reliability. The researcher piloted the instrument with a pool of 120 undergraduate students with the similar characteristics. The quantitative analyses of the collected data were calculated using SPSS software and the findings of exploratory factor analysis were acquired. The factors were extracted. Consequently, the factors with high loadings and factors with no strong loadings on any other factors were deleted. The internal consistency reliability data for the questionnaire was calculated by Cronbach's alpha reliability that produced high reliability coefficients (30 items; α =.906). The final version of the questionnaire with 30 statements that loaded one factor was used to determine the relationship between attitudes towards computer and CBT performance.

3.3. Procedure

In the present research, both quantitative and qualitative instruments were used as the sources to gather research from participants taking both version of CBT and PBT general English test. In the first stage of the study, paper-based version of the test was given to the homogeneous students who were organized into one testing group. This version of the test in which all the question items were presented in four pages is the traditional format of testing that all the undergraduate students are familiar with it. After four weeks interval, testing group took the CBT version of the test in second testing session. The four weeks interval was to reduce the testing, practice and fatigue effects. In CBT session, students could start their exam and scroll through the pages to go forward and backward easily to see and check all question items and options. For a multiple-choice type questions, test taker should read the item and choose one of the four options that they thought was correct or the best answer. Each question had only one option and students could choose only one. If they wanted to change their answer, it was possible simply by marking on the other option then the previous marked option became unmarked. So changing the answers was easy. The system gave the students the option of reminding for each test that they thought they need more consideration (like the tick by pen besides each question that they want back to it later). After answering all test items, students could press the key of finish at the bottom of the page. At the end, the students could see their score by

pressing submit button and expiring the time. At the end of the second testing session, the researcher-made questionnaire was distributed to the students. Then, 20 test takers who filled out the questionnaire were randomly selected and invited for interview. The interviews which were conducted in this study based on an interview guide were applied to confirm the findings from attitude questionnaire. Moreover, the interview questions were prepared in Persian to ease their reporting. The interviewees could use their mother tongue in responding the interview questions to establish a relaxed and comfortable atmosphere among the group.

4. Results and Discussions

To analyze the data, two variables were involved in the study namely test scores on PBT and CBT and attitude towards the use of computer. For analyzing quantitative data, Statistical Package for Social Science (SPSS) software version 21 was used. Qualitative data were analyzed through content analysis. Statistical methods used in this study included paired samples *t*-test and Pearson Correlation analysis to examine the relationship between the attitude towards the computer use and CBT performance.

As one of the main aims of the study was comparing two sets of scores in a within-subject group, a paired-sample *t*-test was run to compare the means of two test scores in order to find out the difference, if any. The probability level for all tests of statistical significance for the study was set at $p < .05$.

Since the parametric statistical analyses of paired-sample t-test and Pearson correlation were supposed to be run to probe the research questions, the researcher had to confirm fulfillment of the four assumptions of interval data, independence of subjects, normality, and homogeneity of variances. The collected data were measured on an interval scale; in addition, the subjects 'performance on the tests was independent of each other and no treatment by peer or group work was administered in this study. The assumption of normality and homogeneity of variances were also met.

Table 1. *Testing Normality Assumption.*

	Tests of Normality					
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	D.F.	Sig.	Statistic	D.F.	Sig.
TG PBT	.185	60	.705	.901	60	.545
TG CBT	.305	60	.816	.747	60	.763

(TG=Testing Group)

Table 1 displays the results obtained from two statistical tests of normality namely Shapiro-Wilks and Kolmogorov-Smirnov. Wilk test is the most powerful test to estimate normal distribution of small sample sizes [70]. From Table 1, p-values were greater than 0.05. Then, it was concluded that the related variables were normally distributed.

Furthermore, Levene's Test of Homogeneity of Variances was run; the result, with an alpha level of .05, p (.504) showed no statistically significance. According to the variances analysis, Levene's F Statistic had a significance value of greater than .05. Then, the assumption of homogeneity of variances was not violated p (.504) >α (.05). It means that the research data had similar variances and using parametric statistical tests was needed to be used.

First, descriptive statistics are used to gain a better view of the data, and then the inferential statistics analysis is displayed to find out the relationship between mean scores. Specifically, means and standard deviations of scores are compared between PBT and CBT versions of test administrations. Table 2 indicates the descriptive statistics of the study.

Table 2. *Descriptive Statistics of Test Scores in both PBT and CBT.*

Descriptive statistics								
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
TG PBT	60	45.4	21.2	3.352	38.61	52.18	26.00	98.00
TG CBT	60	39.6	13.1	2.079	35.39	43.80	20.00	64.00

Of the two versions of the test taken by testing group, the highest mean score was found in PBT, with a relatively higher mean score for PBT than for CBT by 6 points. Based on Table 2, test takers' mean score on PBT (M = 45.4, SD = 21.2) was relatively higher than their mean score on the CBT (M = 39.6, SD = 13.1). On the other hand, the standard deviation in CBT was higher than in CBT which shows that the dispersion of scores from mean score in PBT was higher than in CBT; consequently, it was concluded that Standard Error of Measurement (SEM) in CBT was lower than in PBT.

Since two sets of scores in the present research were received from the same subjects organized into one testing group who took two versions of the same test in two testing sessions, paired t-test formula was run to compare the mean scores of test takers. The main purpose of paired t-test was to collect further evidences to ensure the comparability and interchangeability of two sets of scores. As indicated in Table 3, the t value was 3.36 at $P<0.05$ with 29 degree of freedom. 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=.004).

Table 3. *Paired t-test results for both PBT and CBT modes of administration.*

	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		t	D.F.	Sig. (2-tailed)
				Lower	Upper			
PBT – CBT	2.50	2.73	0.63	0.72	3.36	3.23	59	.004

Then, the results of the paired-sample t-test (t (59) = 3.23, P <.05) indicated that there was a significant difference between the participants' mean scores on CBT and PBT. And based on the descriptive statistics of Table 2, participants performed better on PBT than CBT. Therefore, the results of inferential analysis indicated a statistically significant difference and it was concluded that there was a significant difference in test results of two tests with the priority of PBT over CBT.

Then, a Pearson's product-moment correlation was run to assess the relationship between prior attitudes towards computer and CBT performance of all the test takers of testing group.

Based on the results, while computer attitude had a weak positive correlation with the testing performance, it could not be considered as a dominant factor influencing test scores. The results (r (58) =.216, P >.05) indicated that there was no significant correlation between two test takers' attitudes towards computer use and testing performance variables (Table 4), and computer attitude was not a statistically significant predictor of CBT performance. Thus, the null hypothesis for computer attitude moderator factor was confirmed based on the evidence that the computer attitude was not statistically significant predictor of CBT scores of undergraduate students of PNU.

Table 4. *Pearson Correlation of computer attitude construct with CBT scores of the testing group.*

Correlations			Attitude construct
Spearman's rho	TG CBT	Correlation Coefficient	.216
		Sig. (2-tailed)	.147
		N	58

According to Evans (1996), it was concluded that computer attitude was weakly correlated with changes in test takers' scores in CBT but the difference was not statistically significant. The results showed a positive and weak correspondence between two variables. According to the results, for the testing group, the answers of participants to the second factor and their testing performance was not strongly and positively correlated,.216 (58) =.147, P >.05 [71].

For collecting qualitative data, focus group interview was administered after the second testing session. For focus group interview, 20 participants were selected, 9 of whom preferred computerized test, 7 preferred paper-based test, and 4 favored both testing modes. They were asked the questions to tell their reasons behind their preference to find out the rationales behind each preference to strengthen the finding on quantitative analysis in attitude.

The researcher used the interview guide that helped the interviewer stay on track and keep consistency throughout the interviews with different respondents to ask 7 open-ended questions. Two analysis approaches including quantitative and qualitative ones were employed to analyze the data. The first stage of qualitative analysis of interview data after collecting data was transcription of the recorded voices. In transcription, just the relevant sections of recorded conversations were picked up. Once transcription of the data has been completed, content analysis was conducted by experts in computer and educational testing areas 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 questions were coded and categorized under a common theme [72]. The main relevant and meaningful notions and concepts were identified and categorized under common themes (Table 5).

Table 5. *Results of coding interview Data.*

Preference for CBT	Preference for PBT
Easy to answer	More familiarity with test environment
Easy in changing the answer	Being habituated with test conditions
Speed in test taking	Easier to follow
Novelty	Ordering of questions
Immediacy of reporting the scores	Using strategies in PBT
Preparation for the future exams	Hand noting in PBT
Feeling of progress	No need to extra task demand
More enjoyable	Less risk of technology issues
Lack of human error	
Calm environment	

The finding of interviews as qualitative set of research data were classified into two main themes namely preference for PBT and preference for CBT and the justification for their choices. The results showed higher preference rate for CBT but better performance on PBT.

5. Conclusion

The main purpose of the current study was to discover the equivalency of test scores on PBT and CBT in the English achievement test in Payame Noor University. It also intended to investigate if there was any relationship between computer attitude and test score on CBT. Based upon the quantitative and qualitative data, some major findings were revealed as follows. In general, some major findings can be derived from the current study. Firstly, in comparing the results of computerized and paper-based tests, students showed better performance on PBT than CBT. The results of this study

supports the necessity of doing such comparability studies in higher educational contexts before substituting CBT for PBT or including it in the system [1, 3, 18].

The results of interview supported the quantitative findings, i.e. participants mostly showed high preference for computerized test and liked CBT more than PBT but due to some justifications and habit of taking tests traditionally, they performed better on PBT. This result supports the finding on the relationship between attitude towards the use of computer and test performance on CBT in that despite the high positive attitude for computer using, the respondents performed better on PBT. This result is similar to some previous studies investigating the effect of testing mode preference on CBT and testing performance on it.

The results showed a significant difference between test scores of CBT and PBT with higher score on PBT. The result of this study is in contrast with some other studies demonstrating similar scores between two test administration modes. For example, Anakwe compared the results of computerized testing versus paper-and-pencil testing among 75 undergraduate students each of whom took two computerized tests and two paper-and-pencil tests. The results indicated no difference between two test mode results [15].

The results of this study agree with [1] in finding difference between test scores of PBT and CBT. However, the results of their study from 105 undergraduate students showed higher score on computerized tests compared to paper-based tests. They suggested that the possible reason for higher scores on CBT is due to the nature of the topics covered in the course. Since the title of the course was Computer Fundamentals, they concluded that as the participants were using computers as part of the normal curriculum, it was likely to get better score on computerized test due to the exposure to computers and consequently their more familiarity with computers.

The finding of the present study regarding the difference between test scores on CBT and PBT is in line with [29] that found slightly difference between mean scores with higher performance on PBT. He also argued that this difference could be due to the eye fatigue while reading text through screen as well as the novelty of presenting test through computer in his study context. He also emphasized on the necessity of making students more familiar with such electronic tests by conducting more computerized tests in advance. The difference between this study and [57] study is that he employed English texts to test reading comprehension, whereas in the present study General knowledge of students from their course textbook in the form of multiple-choice test was used. Therefore, the issue of eye fatigue in reading long texts has been dismissed in this study.

Furthermore, it was concluded that there was no significant relationship between attitude and test scores on CBT among the participants of the study. The result of Pearson Correlation analysis showed no significant relationship between these two variables. Although the students showed high positive attitude towards the use of computers, they performed better on PBT. This result could be desirable for the researcher as it is shown that attitude, as a construct irrelevant variable, has no effect on the test performance on CBT.

The results of the present study go with other similar studies such as [29]. He has done a comparability study on CBT and PBT among 167 undergraduate students in Saudi Arabia and took into account the variable of attitude towards the use of computer as a main factor influencing the results on computerized tests. He, similarly, found no significant correlation or interactive effect between the computer attitude and test performance on CBT among his participants while their attitude was positively high towards the use of computer.

The finding of interviews as qualitative set of research data were classified into two main themes namely preference for PBT and preference for CBT and the justification for their choices.

Despite the high percentage of CBT preference reported by the participants, the respondents reported more advantages of PBT over CBT in justifying why they prefer PBT. As it is evident from the results of open-ended question, students mostly preferred CBT than PBT; however, their reasons in focus group interview mostly favored the paper-based tests. The familiarity with the process of doing PBT, the possibility of noting by hand which is one of their test strategies, as well as the similarity of testing with learning materials were among advantages which the advocates of PBT reported in favor of paper-based tests.

The results demonstrated that those who had experienced computer-based test before showed less stress and more preference for computerized tests and some who had not experienced before and favored PBT, changed their choice to CBT due to the reasons they provided in the interview. They, in their justification for changing their choice, mostly stated that the reason of feeling stress before encountering the computer-based test was their more familiarity with taking exams traditionally by paper-and-pencil as well as the habit of taking such tests due to a long time doing it since the primary schools. They then showed their favors for CBT and claimed that changing the old behavior requires time and patience.

On the other hand, as it was mentioned in previous sections, participants preferred to do computerized tests but favored paper-based testing features and performed better on PBT. Computerized test advocators referred to some features of computerized tests that they like more than paper-based test features.

Besides those who favored PBT or CBT, there were other respondents who favored both testing modes and claimed that if one becomes accustomed to and familiar with computerized test gradually, taking test is not different in any type of test administration modes. If the student is ready enough for taking test, doing test in any type looks the same in the condition that computerized tests change to become ordinary like paper-based tests by including it in all stages of education.

Actually, the present limitations of computer analyses of human language did not allow us to address directly the more important assessment of communicative competence. Some other variables such as ethnicity, intelligence, affective and

motivational factors, test anxiety, test effects, test order 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 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.

Since CBT testing programs depend heavily on linear item selection algorithm, the results that will be obtained and presented here will just be specific to CBT testing programs which use this kind of item selection algorithm, and will not be applicable to the other computerized testing programs such as CAT with adaptive algorithm to select and present test items. And, as the next limitation of the current study, it should be mentioned that further research should be done to investigate whether the scoring procedure used in this investigation will produce similar results in testing programs which use other item selection algorithm.

It is hoped that this study add to the value of the current comparability research that focus on computer-based-tests. Decision makers, program developers, and educators are suggested to include computerized exams in lower level of education from elementary school. The results is useful for language institutes as well because using computer in teaching and testing English language simultaneously enhance their students' sense of being developed by technology and hence they use new strategy in learning language through computers which in turn may enhance their language learning. Consequently, they can make their English learners more ready for doing computerized and online TOEFL, IELTS, GRE, or similar determinant exams.

References

[1] 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 (5) 593-602. https://doi.org/10.1111/1467-8535.002944.

[2] Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*. 17 (1), 1–42. https://doi.org/10.1191/026553200675041464 4.

[3] Hashemi Toroujeni, S. M. (2016). *Computer-Based Language Testing versus Paper-and-Pencil Testing: Comparing Mode Effects of Two Versions of General English Vocabulary Test on Chabahar Maritime University ESP Students' Performance*. Unpublished thesis submitted for the degree of Master of Arts in TEFL. Chabahar Marine and Maritime University (Iran) (2016).

[4] Peat, M., & Franklin, S. (2002). Supporting student learning: the use of computer based formative assessment modules. *British Journal of Educational Technology*, 33 (5), 515–523. https://doi.org/10.1111/1467-8535.002888.

[5] Bennett, R. E. (1998). *Reinventing assessment: Speculations on the future of large scale educational testing*. Princeton, NJ: Educational Testing Service, Policy Information Center.

[6] Bunderson, V., Inouye, D. & Olsen, J. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed). *Educational Measurement*, pp. 367-407. Phoenix, AZ: Oryx Press.

[7] Chapelle, C. (2007). Technology and second language acquisition. Annual Review of Applied Linguistics. 27, 98-114. https://doi.org/10.1017/s0267190508070050.

[8] Mazzeo, J., & Harvey, L. A. (1988). The equivalence of scores from automated and conventional education and psychological tests: a review of the literature. (Report No. CBR 87-8, ETS RR 88-21). Princeton, NJ: Educational Testing Services.

[9] Mead, A. and Drasgow, F. (1993). Equivalence of Computerized and Paper-and- Pencil Cognitive Ability Tests: A Meta-Analysis. Psychological Bulletin, 114 (3), pp. 449-58. https://doi.org/10.1037/0033-2909.114.3.449.

[10] Wainer, H., Doran, N., Flaugher, R., Green, B., Mislevy, R., Steinberg, L. & Thissen, D. (1990). Computer Adaptive Testing: A Primer. Hillsdale. NJ: Lawrence Erlbaum Associates.

[11] Higgins, J., Russell, M. & Hoffmann, T. (2005). Examining the Effect of Computer-Based Passage Presentation on Reading Test Performance. The Journal of Technology, Learning, and Assessment, 3 (4), pp. 5-34.

[12] Khoshsima, H., Hosseini, M. & Hashemi Toroujeni, S. M. (2017). *Cross-Mode Comparability of Computer-Based Testing (CBT) versus Paper and Pencil-Based Testing (PPT): An Investigation of Testing Administration Mode among Iranian Intermediate EFL learners.* English Language Teaching, Vol. 10, No. 2; January (2017). ISSN 1916-4742 (Print), ISSN (1916-4750). http://dx.doi.org/10.5539/elt.v10n2p23.

[13] Russo, A. (2002). Mixing Technology and Testing, Computer-Based Testing, The School administrator, http://www.aasa.org/SchoolAdministratorArticle.aspx?id=10354, 9.05.2012.

[14] Trotter, A. (2001). Testing firms see future market in online assessment. Education Week on the Web, 20 (4), 6.

[15] Anakwe, B. (2008). Comparison of student performance in paper-based versus Computer-based testing. Journal of Education for Business, September/October, 13-17. https://doi.org/10.3200/JOEB.84.1.13-17.

[16] Chapelle, C. A. & Douglas, D. (2006). *Assessing language through computer technology*. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511733116.

[17] Paek, P. (2005). Recent trends in comparability studies. Pearson Educational Measurement Research Reports. Research Report 05-05. Pearson Educational Measurement. USA.

[18] Khoshsima, H. & Hashemi Toroujeni, S. M. (2017c). *Technology in Education: Pros and Cons of Using Computer in Testing Domain.* International Journal of Language Learning and Applied Linguistics World (IJLLALW), (2017) Volume 1 (2), February 2017; 32-49. EISSN 2289-2737, PISSN: 2289-3245. http://ijllalw.org/Current-Issue.htmll.

[19] Douglas, D. (2000). Assessing Languages for Specific Purposes. Cambridge. Cambridge University Press.

[20] Fazeli, P. L., Ross, L. A., Vance, D. E., & Ball, K., (2013). The relationship between computer experience and computerized cognitive test performance among older adults. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences,* 68 (3), 337–346. https://doi.org/10.1093/geronb/gbs071.

[21] Khoshsima, H. & Hashemi Toroujeni, S. M. (2017a). *Transitioning to an Alternative Assessment: Computer-Based Testing and Key Factors related to Testing Mode.* European Journal of English Language Teaching, Vol. 2, Issue. 1, pp. 54-74, February (2017). ISSN 2501-7136. http://dx.doi.org/10.5281/zenodo.2685766.

[22] Lightstone, K., Smith, S. M., (2009). Student Choice between Computer and Traditional Paper-and-Pencil University Tests: What Predicts Preference and Performance? International Journal of Technologies in Higher Education, 6 (1), 30-45. https://doi.org/10.7202/039179ar.

[23] Maguire, K. A., Smith, D. A., Brallier, S. A., & Palm, L. J. (2010). Computer-Based Testing: A Comparison of Computer-Based and Paper-and-Pencil Assessment. Academy of Educational Leadership, 14 (4), 117-125.

[24] Terzis, V., & Economids, A. A. (2011). Computer based assessment: Gender differences in perceptions and acceptance, *Computers in Human Behavior* 27 (2011), 2108-2122. https://doi.org/10.1016/j.chb.2011.06.005.

[25] Bennett, R. E. (2002). Inexorable and inevitable: the continuing story of technology and assessment. The Journal of technology, Learning, and Assessment, 1 (1), 23.

[26] Dooling, J. (2000). What students want to learn about computers? Educational Leadership, 58 (2), 20-24.

[27] 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).

[28] Sawaki, Y. (2001). Comparability of Conventional and Computerized Tests of Reading in a Second Language. Language Learning & Technology. 5 (2), 38-59.

[29] 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).

[30] Khoshsima, H. & Hashemi Toroujeni, S. M. (2017b). *Comparability of Computer-Based Testing and Paper-Based Testing: Testing Mode Effect, Testing Mode Order, Computer Attitudes and Testing Mode Preference.* International Journal of Computer (IJC), (2017) Volume 24, No 1, pp 80-99. ISSN 2307-4523 (Print & Online), http://ijcjournal.org/index.php/InternationalJournalOfComputer/article/view/825/41888.

[31] Wise, S., & Plake, B. (1990). Computer-Based Testing in Higher Education. *Measurement and Evaluation in Counselling and Development*, 23, 10.

[32] Choi, I., Kim, K., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. Language Testing, 20 (3), 295-320.

[33] Neumann, G. & Baydoun, R., (1998). Computerization of paper-and-pencil tests: When are they equivalent? Applied Psychological Measurement, 22, 71–83. https://doi.org/10.1177/01466216980221006.

[34] Bodmann, S. M. & Robinson, D. H. (2004). Speed and Performance Differences among computer-based and paper-based tests. Journal of Educational Computing Research, 31 (1) 51-60. https://doi.org/10.2190/GRQQ-YT0F-7LKB-F033.

[35] Eid, G. K. (2005). An investigation into the effects and factors influencing computer- based online math problem-solving in primary schools. Journal of Educational Technology Systems, 33 (3), 223-240. https://doi.org/10.2190/J3Q5-BAA5-2L62-AEY3.

[36] Puhan, G., Bought on, K., & Kim, S. (2007). Examining Differences in Examinee Performance in Paper and Pencil and Computerized Testing. *The Journal of Technology, Learning, and Assessment*, 6 (3), 5-20.

[37] Kapes, J. T., Matinez, L., & Ip, C. F. (1998). Internet-based vs. paper-pencil occupational competency test administration: an equivalency study. Journal of Vocational Education Research, 23 (3), 201-219.

[38] Choi, S. W. & Tinkler, T. (2002). Evaluating comparability of paper and computer-based assessment in a K–12 setting. *Paper presented at annual meeting of the National Council on Measurement in education*, New Orleans, LA.

[39] 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.

[40] O'Malley, K. J., Kirkpatrick, R., Sherwood, W., Burdick, H. J., Hsieh, C. & Sanford, E. E. (2005). 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.

[41] Pomplun, M., & Custer, M. (2005). The score comparability of computerized and paper-and-pencil formats for K-3 reading tests. Journal of Educational Computing Research, 32 (2), 153-166. https://doi.org/10.2190/D2HU-PVAW-BR9Y-J1CL.

[42] Mason, B. J., Patry, M., & Bernstein, D. J. (2001). An examination of the equivalence between non-adaptive computer-based and traditional testing. Journal of Educational computing research. 24, 29-39. https://doi.org/10.2190/9EPM-B14R-XQWT-WVNL.

[43] Kruger, J., Wirtz, D., & Miller, D. T. (1977). Counterfactual thinking and the first instinct fallacy. *Journal of Personality and Social Psychology.*

[44] Schwarz, S. P., McMorris, R. F., & DeMers, L. P. (1991). Reasons for changing answers: An evaluation using personal interviews. Journal of Educational Measurement, 28, 163-171. https://doi.org/10.1111/j.1745-3984.1991.tb00351.x.

[45] Vispoel, W. P. (1998). Reviewing and changing answers on computer-adaptive and self-adaptive vocabulary tests. Journal of Educational Measurement, 35, 328-345. https://doi.org/10.1111/j.1745-3984.1998.tb00542.x.

[46] Parshall, C. G., & Kromery, J. D. (1993). Computer versus paper and pencil testing: An analysis of examinee characteristics associated with mode effect. *Abstract from: ERIC Abstract No. ED363272.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

[47] Johnson, M., and Green, S. (2006). On-line mathematics assessment: The impact of mode on performance and question answering strategies. The Journal of Technology, Learning, and Assessment, 4 (5).

[48] Hargreaves, M., Shorrocks-Taylor, D., Swinnerton, B., Tait, K., & Threlfall, J. (2004). Computer or paper? that is the question: Does the medium in which assessment question are presented affect children‘s performance in mathematics? Educational Research, 46 (1), 29-42. https://doi.org/10.1080/0013188042000178809.

[49] Horkay, N., Bennett, R. E., Allen, N., & Kaplan, B. (2005). Online assessment in writing. In B. Sandene, N. Horkay, R. E. Bennett, N. Allen, J. Braswell, B. Kaplan, & A. Oranje (Eds.), Online assessment in mathematics and writing: Reports from the NAEP Technology-Based Assessment Project (NCES 2005-457). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

[50] Sandene, B., Horkay, N., Bennett, R. E., Allen, N. Braswell, J., Kaplan, B., & Oranje, A. (Eds.) (2005). *Online assessment in mathematics and writing: Reports from the NAEP technology based assessment project* (NCES 2005-457). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

[51] Fletcher, P., & Collins, M. A. J. (1986). Computer-administered versus written tests-advantages and disadvantages. *Journal of Computers in Mathematics and science Teaching*, 6, 38-43.

[52] Messick, S. (1989). Validity. In Robert Linn (Eds) *Educational Measurement* (3rd Ed.) London: Collier Macmillan Publishers.

[53] Chapelle, C. (2001). *Computer Applications in Second Language Acquisition: Foundations for teaching, Testing and Research*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524681.

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

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

[56] McNamara, T. (2000). *Language Testing*. Oxford: Oxford University Press.

[57] Al-Amri, C (2008). Computer-Based Testing vs. Paper-Based Testing: A Comprehensive Approach to Examining the Comparability of Testing Modes. Essex Graduate Student Papers in Language & Linguistics, 10, 22-44.

[58] Tatira, B., Mutambara, L. H. N., Chagwiza, C. J., & Nyaumwe, L. J., (2011). Computerized Summative Assessment of Multiple-choice Questions: Exploring Possibilities with the Zimbabwe School Examination Council Grade 7 Assessments. Computer and Information Science. 4 (6). https://doi.org/10.5539/cis.v4n6p66.

[59] 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 (2), 219-274. https://doi.org/10.1111/0023-8333.00088.

[60] Zhang, Q. (2007). EFL Teachers’ Attitudes toward Information and Communication Technologies and Attributing Factors. Peking University.

[61] 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.

[62] Goldberg, A., & Pedulla, J. J. (2002). Performance differences according to test mode and computer familiarity on a practice GRE. Educational and Psychological Measurement, 62 (6), 1053-1067. https://doi.org/10.1177/0013164402238092.

[63] Pomplun M., Ritchie, T., & Custer M. (2006). Factors in paper-and-pencil and computer reading score differences at the primary grades. *Educational Assessment*, 11 (2), 127-143. https://doi.org/10.1207/s15326977ea1102_3.

[64] Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment,* 6 (9).

[65] Kutluca, T., & Gokalp, Z., (2011). A study on computer usage and attitude toward computers of prospective preschool teachers. *International Journal on New Trends in Education and Their Implications,* 2 (1), 2-147.

[66] Rezaee, A. A., Abidin, Z. M. J., Isa, H. J., & Mustafa, O. P. (2012). TESOL in-Service Teachers' Attitudes towards Computer Use. English Language Teaching. 5 (1), 61-68.

[67] Powers, D. and O‘Neill. K. (1993). Inexperienced and anxious computer users: coping with a computer-administered test of academic skills. *Educational Testing Services, Research Report* RR 92-75. https://doi.org/10.1207/s15326977ea0102_4.

[68] Warner, R. M. (2013). Applied Statistics: From Bivariate through Multivariate Techniques. (2th Ed.). SUA: SAGE Publication Inc.

[69] Privitera, G. J. (2012). Statistics for the Behavioral Siences. USA: SAGE publication Inc.

[70] Ricci, V. (2005). *Fitting distributions with R. R project.* Website http://cran.r-project.org/doc/ contrib/Ricci-distributions-en.pdf. Retrieved July 6, 2007.

[71] Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole Publishing.

[72] Seidman, I. (1998). *Interviewing as qualitative research: A guide for researchers in education and the social sciences* (2nd ed.). New York: Teachers College Press.