Score Equivalence, Gender Difference, and Testing Mode Preference in a Comparative Study between Computer-Based Testing and Paper-Based Testing

https://doi.org/10.3991/ijet.v14i07.10175

Mohammad Reza Ebrahimi
Gonabad University, Gonabad, Iran

Seyyed Morteza Hashemi Toroujeni(✉),Vahide Shahbazi
Chabahar Maritime University, Chabahar, Iran
hashemi.seyyedmorteza@gmail.com

Abstract—Score equivalency of two Computer-Based Testing (henceforth CBT) and Paper-and-Pencil-Based Testing (henceforth PBT) versions has turned into a controversial issue during the last decade in Iran. The comparability of mean scores obtained from two CBT and PBT formats of test should be investigated to see if test takers' testing performance is influenced by the effects of testing administration mode. This research was conducted to examine score equivalency across modes as well as the relationship of gender, and testing mode preference with test takers' performance on computerized testing. The information on testing mode preference and attitudes towards CBT and its features was supported by a focus group interview. Findings indicated that the scores of test takers were not different in both modes and there was no statistically significant relationship between moderator above variables and CBT performance.

Keywords—Computer-based testing, paper-based testing, score equivalence, testing mode Preference

1 Introduction

Sometimes technological developments have such great influences on human life that some scholars and sociologists categorize mankind history based on the produced technological tools. Our lifestyle has been considerably changed by technology; it has exerted a great impact on professions, thinking, communication, and also all aspects of our lives have been influenced by it. [1]. For example, technology is being used to provide students with useful information to create and connect learning groups to create a convenient learning environment [2]. According to the assessment researcher, Stuart Bennett- a quite committed and enthusiastic proponent of technology- who is interested in researching measurement, new technology's transformative impacts on assessment domain makes it possible to impel someone manages something well and satisfactorily by building some tests based on the conceptualization of preconditions and qualifications. He also declared that by enjoying technological assessment tools to

create tests, test takers' performance could be practically assessed through computer-based simulations, item, and item bank creation and also scoring process. Besides, the large-scale delivery test is made possible by enjoying technology and computer in the assessment domain [3]. The first talks about the new technology's transformative influences on assessment domain that were also mentioned by Bennett have been organized much earlier. It is highly likely that teachers utilize computers in administering tests as they are readily available [4].

CBT turned to a controversial research area as how to develop and administer high stakes computerized version of testing program in 70s A.D.; however it must be noted that ASVAB (Armed Services Vocational Aptitude Battery program done by USA Defense Department, the Graduate Record Examination (GRE), Test of English as a Foreign Language (TOEFL) and etc.) the real history of computerized fixed testing goes back to the decade of 30s A.D. dates back to 30s A.D. For the first time, the IBM model 805 machine was used in 1935.

It aimed at scoring objective tests of millions of American test takers each year. Use of computer in language testing has resulted in the birth of independent discipline named CBT (Computer-Based Testing) which has been expedited by CAL (Computer-Assisted Learning). CBT has changed the nature of the language assessment field with its potential benefits and capabilities. CBT may assist language assessment field by helping overcome many common administrative and logistical problems that are widespread in the traditional fixed-length testing environment. By offering new approaches and basic advantages such as easier and more precise test scoring and reporting, item innovation, item generation, greater security, standardization, and test efficiency, test booklets and answer sheet elimination, more flexible scheduling, reduced measurement errors, and etc., CBT opened new windows and laid foundations for future assessment in educational testing. Due to the paradigm differences in test delivery that range from linear or fixed to adaptive test delivery, CBT has been employed to refer to the fixed-length, fixed form computerized kind of a test without any adaptive nature of item selection seen in adaptive testing.

Bit by bit, for any fixed form computerized exam on any content Fixed-length kind of test delivery (i.e., Computer-Based Testing or CBT) has started to be utilized; thus, in the current investigation, CBT initialism was used to refer to test delivery of language content named after Computerized Fixed Language Test and to recognize computerized test of language content from the other kinds of computerized tests of any other contents.

The proposed CBT by this research meets the most advantages of computerized test administration among them presenting items on the screen, faster and easier test scoring and result reporting by a computer, greater security, revising the answers that are not allowed in most adaptive tests, and item innovation such as audio and video prompt can be mentioned. CBT characterized by a direct and uncomplicated scoring algorithm and quickly changing content (e.g., vocabulary to grammar, or grammar to any other content) has been considered an acceptable delivery method and is going to be typically applied to the small-scale educational achievement test delivered in educational contexts in Iran.

The issue that currently needs more attention and prompt investigation of researchers is to study the testing mode effects on comparability and equivalency of the data obtained from two modes of presentation, i.e., traditional paper-and-pencil (PBP) and computerized tests. According to [5], comparability researches and studies in second language tests are in short supply, and he also emphasized over the importance of conducting comparability studies in local settings to detect any potential test-delivery-medium effect especially when a traditional PBT test is converted to a computerized one.

The critical issue of establishing comparability and equivalency of computerized test with its paper-and-pencil counterpart is of prime importance. Some researchers have focused on the equivalency of computer and paper-administered tests in terms of scores [6]. Recently, some studies have been conducted to indicate that to replace the computer-based test with conventional paper-and-pencil one; we need to prove that these two versions of the test are comparable. In other words, the validity and reliability of computerized counterpart are not violated. Actually, there is no agreed-upon theoretical explanation for the test mode effects. The comparability is achieved through equivalent scores of two test versions.

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. Such developments and widespread access to a computer, especially in educational contexts, have greatly influenced many areas of interests and subjects [7] such as the English testing domain. In addition, this is the reason that some international macro organizations dealing with conducting TOEFL, IELTS, GRE, etc. started to give their offline or online examinations in the computerized version. Implementation of these tests through computers by the testing organizations is in the direction of findings of several studies in which a high level of agreement and acceptance to use computer-based tests is revealed [8].

Since in Iran, however, computerized testing is still at an early experimental stage, the present study would be conducted to provides some helpful and informative findings for those learners, teachers, testing practitioners and researchers who seek to know the possibility of replacing computerized tests with paper-pencil ones. It is done to show the comparability between two test modes of administration. 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 the two test versions. It means that whether there is any discrepancy that violates the reliability and validity of the computerized counterpart; the computerized version that is supposed to be replaced with the conventional paper-and-pencil version of the test. In the case of [6], significant cross-mode differences in mean of listening, grammar, and vocabulary subtests were observed. In this study, the largest cross mode discrepancy was observed in the reading comprehension subtest. But they explained that the indicators of incomparability seen in the results might be due to the discrepancy in various test layouts across test presentations rather than the content of the tests itself.

The results of these studies have substantially influenced current approaches to investigate comparability between two versions of a test. Though such tests are not prevalent and popular in Iran because of test takers' unfamiliarity with such kind of

exams, it is necessary to make individuals engaged in learning settings familiar with computerized tests. Familiarity with computer and testing settings in which computers are used as a medium to present items is so critical that some test takers asserted that their computerized test results are not representative of their proficiency level due to their lack of familiarity with computerized testing.

Following the researchers who have done studies to investigate the effects of using computers in testing and assessment and examining the administration mode effects of testing, we can observe some advantages of using computers as the most effective technological tool in testing and assessment field. Some prerequisites should be met to enjoy the positive effects of computers in the testing field. In this direction, [9] stated that to establish a valid and reliable computerized test and to replace it with its paper-and-pencil counterpart, equivalent test scores of two versions should be established. It is exactly what the comparability of CBT and PBT means.

To elucidate the concept of comparability, it should be stated that the basis of linear computer-based and paper-based testing is a test theory called classical true-score test theory that supports those set of testing standards that have to be observed within computer-based testing [10]. According to this theory, two sets of almost similar test scores should be received by test takers who are involved in the same test with two different administration modes.

The standards for developing computerized testing to administer and replace with its paper-and-pencil counterpart require that equivalent test scores be established for PBT and CBT modes. Although two testing modes have been nearly identical in some comparability studies, significant discrepancies of test scores can also be observed in some other ones. Therefore, the validity of replacing CBT with its PBT counterpart in educational assessment in academic contexts has often been under question. Then, as the first step to replace a CBT program with its PBT counterpart in Chabahar Maritime University of Iran, a comparability study comparing testing mode effects of two versions of general English Vocabulary Test on test takers' performance will be done to see whether the two sets of scores are comparable and consequently valid or reliable. And it is important to see whether test results received from the CBT version have the same features as the scores derived from the linear PPT version.

Converting the conventional PPBT version of a test into its computerized counterpart might become problematic when considering reliability and validity. Creating reliable and valid tests is the main issues and concerns in utilizing CFLT. Johnson and Green state that just a CFLT that is matched with its counterpart's validity and reliability can assist the test takers in fulfilling their needs [11]. The evaluation of validity and reliability issues is, therefore, the reason for doing many of the comparability studies between CFLT and PPBT [12] [13]. A test is reliable when it regularly measures what it is expected to measure by producing stable and constant scores on two testing occasions. In other words, a test can be considered reliable when constant similar results or scores are repeated under the same conditions [14]. For example, a vocabulary test that gives three various marks on three successive occasions without applying any change to the test would not be a reliable test. According to Bachman and Palmer, the degree to which a test produces reproductive and consistent results is defined as reliability [15]. Therefore, it is important to examine the reliability and validity of a computer-

ized test by conducting a comparability study, particularly, in a local contextto establish any testing mode effects that result from converting a conventional test into its computerized counterpart. The same scores obtained from two versions of the same test demonstrate that the test is reliable. One of the major goals pursued in comparability studies is to examine the interchangeability of test scores across different testing modes of administration. To achieve this goal, test items should be presented uniformly across two modes.

However, we can expect the same or evenly matched scores in both modes of administration when we administer two equivalent tests or two versions of the same test covering similar materials; the more identical and interchangeable the findings of two similar or equivalent tests are, the more reliable the test is [16] [17]. When tasks are moved from pen and paper to the computer, equivalence is often assumed, but this is not necessarily the case. For example, even if the paper version is valid and reliable, the computer version may not exhibit similar characteristics. If equivalence is required, then this needs to be established [18].

Hence, in brief, the main objective of the current investigation is developing a computerized version of English Vocabulary in Use test published by Cambridge University Press and to calculate the comparability and interchangeability of the PBT and CBT versions of the same test and furthermore, in spite of the existence of mode differences between them to check if the two versions of the test are equivalent.

The second and final aim is to study how gender difference and testing mode preference might affect test scores and test takers' performance (to what extent do these moderator factors moderate the effects of administration mode on the scores magnitude from the tests).

2 Literature Review

Employing computer-based testing is rapidly growing for several reasons [19]. Two identical paper-based and computer-based tests may not necessarily provide the same results; such empirical findings which help testing practitioners decide whether to replace computer-based testing with paper-based testing are referred to as "Testing Mode Effect." However, the researchers have not yet reached an agreement on a comprehensive theoretical explanation for testing mode effect. In several testing mode effect studies, although the content and the cognitive activity of two paper-based and computerized tests are identical, significant differences are usually observed between two sets of achieved scores. Bunderson and the colleagues reported the superiority of CBT over PBT in three studies [20]; Also, Khoshsima and Hashemi Toroujeni indicated the priority of CBT over PBT [21].

On the other hand, some other studies reported no statistically significant difference between paper-based and computerized tests [22].

Moreover, studying the testing mode effects on the equivalency of data gained from two different presentation modes, namely conventional PBT and CBT calls for more investigation. Even two identical tests or similar versions of the same test would not produce identical results due to some diversities including exact content on two versions of the same test, environmental variables such as fatigue in conventional paper-

based test or light of screen in onscreen tests, as well as students' error in responding. Even the same test that is administered in a day, to the same group of test takers may result in different sets of scores that do not coincide with each other in the other day due to the variables above. However, when two tests or the same test of two versions covering the similar materials are administered, one would prefer that students' scores be similar in both. The more comparable the scores are, the more reliable the test scores will be [23] [24].

Availability of computerized form of standardized tests provides users with the choice of taking the test in whichever mode. Converting paper and pencil assessment tools into computerized versions often requires that the computerized form be comparable to the conventional paper and pencil one and the scores and the results obtained from two identical test forms approximate to each other. Interchangeability is required when students may take the same test in either mode [25].

Converting PBT into CBT should be done through carefully well-organized empirical investigations. The empirical investigations examine the existence of distinctive effects caused by changing administration mode from conventional PBT to modern CBT. Conducting these kinds of comparability investigations help test practitioners to see if the scores obtained from computerized tests remain valid and that students are not disadvantaged by taking CBT.

Testing mode preference of test takers that are typically related to high stakes standardized test administration is being noticed in recent research. Like this study, some others have been done to examine the preference of test takers on testing administration mode [26] [27] [22] [28]. The researcher of the current study investigated the influence of test takers' preference on their test performance on CBT by employing questionnaire and interviews. While in several studies, the effectiveness difference of methods regarding race, age, and gender was examined, and no statistically significant difference in their actual performance was found [29], in some other comparability studies such as [30] [21], statistically significant difference was found. Additionally, [31] investigated the relationship of gender with CBT performance and the trends of female and male test takers towards the features of CBT.

Considering both theoretical and pedagogical perspectives, the following research questions were addressed to achieve the research objectives:

- Is there any statistically significant difference between the performance of computer-based testing and paper-and-pencil-based testing?
- Do participants' gender difference and prior testing mode preferences affect their performance on CBT?
- Do participants perform better on their preferred test mode?

3 Methodology

3.1 Research design

A mixed-methods approach which combined multiple-choice achievement tests, questionnaires and interviews was the methodological approach that was employed in this study. As the first critical step, this comparability study used a common person design to organize a testing group which is a powerful design in detecting differences especially in a smaller sample of test takers to collect good data for making score comparison. Participants were assigned to one testing group in which the testing mode of administration that was considered the treatment in this study was investigated.

3.2 Participants

The 120 intermediate graduate students as test takers of the research whose English proficiency level was intermediate were selected from those 165 homogenous students who took a placement test. The number of female students (n=57%) exceeds the number of male participants (n=43%). The age range of all the 120 students was between 22 to 26 years. Mean age was 23.5 years with a standard deviation of 2.51.

3.3 Instruments

Nelson Proficiency Test (test 150 C) which was selected from Nelson English Language Tests by Fowler and Coe [32] was the first research data collection instrument that was distributed to 263 graduate students of CMU to determine their proficiency level. After implementing the placement test to determine the homogeneity of test takers, paper and pencil version of English Vocabulary in Use Pre-intermediate and Intermediate Level Test was administered at the end of the course teaching period. A professional web-based testing service provided by Classmarker.com website was used to administer the CBT version of the test.

Nowadays, teaching and learning processes are becoming updated using the internet, and teacher-centered education is being substituted by learner-centered education [33].

Test takers were required to read a question on the computer screen and choose the most appropriate option under each question by clicking the mouse on the blank space next to the options. Like the PBT version of the test, test takers could review and change their answers by changing the tick from one selected option to another one. They could even go back to the previous page to review and change their answers.

Another instrument to collect the research data concerning to the second research question was a simple question mentioned at the bottom of exam paper and screen, i.e., would you prefer taking a test on paper – no difference – computer to examine the relationship between testing mode preference and performance. The feelings and impressions of test takers about CBT were studied after their exposure to CBT by a researcher-made simple questionnaire. This instrument that is a set of researcher-made questions regarding the testing mode preference assessed the development of positive

or even negative attitudes towards CBT. And 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 participants were asked about their attitudes towards the features of two modes of testing administration, testing mode preference, development of positive or even negative attitudes and their reasons for possible changing mode preference.

4 Results and Discussion

First of all, the internal consistency for both paper and computer-based tests were calculated, and relatively high-reliability coefficients (for PBT, α=93 & for CBT, α=95) were achieved. Shapiro-Wilks and Kolmogorov-Smirnov statistical tests were used to provide objective judgment of normality rather than skewness and kurtosis. According to the results and given p=.752 for PBT version and p= 819 for the CBT version, it was concluded that each of the levels of the independent variable was normally distributed. Therefore, the assumption of normality was met for this study.

Furthermore, based on the results of Levene's Test, F (1,239) =7.7, p=.0, with an alpha level of .05, p (.697), the assumption of homogeneity of variances is satisfied, p (.697) > α (.05). It means that our data had similar variances and we can use parametric statistical tests.

Of the two formats of the test taken by the testing group, the highest mean score was found in PBT, with a relatively higher mean score by .53 points. Test takers' mean score on PBT (M = 46.66, SD=17.43) was a little bit higher than their mean score on the CBT (M=46.13, SD=13.8). On the other hand, the standard deviation in PBT was higher than in CBT. It meant that the dispersion of scores from the 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.

According to the findings of the One-Way ANOVA test (Table 1), there was not any statistically significant difference in scores between PBT and CBT at a .05 level. Based on the results of the score analysis of two testing sessions, the Sig. value was .896 at P<0.05. This amount of significance value at 119 (N-1) degree of freedom in a .05 level revealed that there was no significant difference between two sets of scores obtained from two formats of the test and the test scores of participants were not different in paper-based and computer-based versions of the test (Sig=.896, P>0.05).

Table 1. One-way ANOVA comparing scores of participants in PBT & CBT

ANOVA					
	Sum of Squares	***D.F.***	***Mean Square***	***F***	***Sig.***
Between Groups	4.267	1	4.267	.017	.896
Within Groups	14338.133	238	247.209		
Total	14342.400	239			

Additionally, based on the results, male participants' mean score on CBT (M=45.66, SD=14.98) was higher than female participants' mean score on CBT (M=44.66, 5.46) (Table 2).

Of the male and female CBT sessions, the highest mean score was found in male CBT, with a relatively higher mean score by 1 point. On the other hand, the standard deviation in male CBT was higher than in female CBT. It means that the dispersion of scores from the mean score in male CBT was higher than in female CBT; consequently, it was concluded that Standard Error of Measurement (SEM) in female CBT was lower than in male CBT. According to the results of the analysis on male participants' scores on CBT and female participants' scores on CBT, the Sig observed value was .875 at $P<0.05$.

Table 2. Distribution of Female Participants' CBT Scores versus Male Participants' CBT Scores

Descriptive Statistics								
	N	***Mean***	***Std. Deviation***	***Std. Error***	***95% Interval for Lower Bound***	***Confidence Mean Upper Bound***	***Minimum***	***Maximum***
Male	2	45.6	14.98	3.05	39.338	51.994	20.00	64.00
CBT	4				6	7		
Female	6	44.66	5.46	2.23	38.931	50.401	38.00	50.00
CBT					5	9		
Total	3	45.46	13.54	2.47	40.40	50.523	20.00	64.00
	0				4	9		

Table 3. One-way ANOVA comparing CBT scores of female participants versus CBT scores of male participants

ANOVA					
	Sum of Squares	***D.F.***	***Mean Square***	***F***	***Sig.***
Between Groups	4.800	1	4.800	.025	.875
Within Groups	5314.667	28	189.810		
Total	5319.467	29			

Therefore, one way ANOVA analysis showed that the differences between the male participants' scores in CBT version (n=68, M=45.66, SD=14.98) and female participants scores in CBT version of the test (n=52, M=44.66, SD=5.46) were not statistically significant, Sig=.875, $p>0.05$.

To answer research question two, responses to the simple question appeared at the bottom of PBT version of testing group one were correlated with participants' mean score on the computerized test to see if there was any significant correlation between their prior testing mode preference and testing performance on CBT. The researcher also performed multiple comparisons between different preference groups using descriptive statistics 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 the testing group. There was a weak negative correlation between both pre and post-CBT mode preference and CBT performance of

the testing group, r (118) =.-.017, p < .918 and r (118) =.-.112, p < .490, respectively (Table 4).

Table 4. Pearson correlation of Pre-CBT and Post-CBT mode preference with CBT scores of testing group

Pearson Correlations		**Pre-CBT Mode Preference**	**Post-CBT Mode Preference**
Testing Group CBT Performance	Pearson Correlation	-.017	-.112
	Sig. (2-tailed)	.918	.490
	N	120	120

In the next step, descriptive statistics of different testing mode preference groups of the testing group were used to gain a better view of the data. The descriptive statistics output is displayed in Table 5.

Table 5. . PPT Performance of different preference groups of testing roup

Pre-CBT Mode Preference	**N**	**PPT Mean Score**	**Std. Deviation**	**Std. Error**	**95% Confidence Interval for Mean**		**Minimum**	**Maximum**
					Lower Bound	***Upper Bound***		
On Paper	85	40.57	8.34	1.57	37.3368	43.8061	28.00	52.00
No Difference	15	47	1.06	.377	46.1063	47.8937	46.00	48.00
On Computer	20	64	.00	.000	64.0000	64.0000	64.00	64.00
Total	120	44.20	9.98	1.57	41.0074	47.3926	28.00	64.00

As shown in Table 5, the PBT mean score of On-Computer preference group (PBT/M=64, (SD=.0)) was higher than the other two preference groups. It means that the persons who preferred CBT over PBT did better than those who preferred PBT (PBT/M=40.57, (SD= 8.34)) on the PBT version of the test. On the other hand, the persons who preferred CBT did better than the other preference groups on the PBT version. On the other hand, those who preferred taking the PBT version of the test in the PBT testing session had better performance on the CBT testing session (CBT/M=41.42, (SD=15.95)). But, the test takers who preferred taking the test in the CBT version did not perform better in their preferred testing mode.

To compare the results of different testing mode preference groups of testing group on PBT and CBT sessions, as Table 5 revealed, those participants who preferred taking PBT version of the test (PBT/M=40.57, (SD=8.34)) outperformed in their CBT exam (CBT/M=41.42, (SD=15.95)) (Table 6). Accordingly, those who preferred taking the test on CBT (PBT/M=64, (SD=0)) (Table 5), did the same on their CBT exam (CBT/M=64, (SD=0)). And those who didn't mind taking the test on either mode (PBT/ M = 47, (SD = 1.06)), did better on CBT (CBT/M=56, (SD=4.27)) (Table 6). However, the overall results of prior testing mode preference and testing performance of different preference groups' analysis indicated that there was not necessarily positive interaction between testing mode preference and testing performance. The reason might be either the testing orders, i.e., administration of CBT in the first testing session for testing group two or the novelty of CBT in the target setting [12] [22].

Table 6. CBT Performance of different preference groups of testing group

Pre-CBT Mode Preference	N	CBT Mean Score	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	*Upper Bound*		
On Paper	85	41.42	15.95	3.01	35.2429	47.6142	14.00	66.00
No Difference	15	56	4.27	1.51	52.4250	59.5750	52.00	60.00
On Computer	20	64	.00	.00	64.0000	64.0000	64.00	64.00
Total	120	46.6	15.74	2.48	41.5652	51.6348	14.00	66.00

The CBT mean score of On-Computer preference group (CBT/M=64, (SD=.0)) was higher than the other two preference groups, regarding the performance of different preference groups of testing group on CBT version of the test (the second version that they took in their second testing session), and it was shown in Table 6, in the second testing session of testing group, i.e., CBT version.

It means that the persons who preferred CBT over PBT did better than those who preferred PBT (CBT/M=41.42, (SD=15.95)) on the CBT version of the test. On the other hand, those who didn't mind taking the test on either mode (PBT/ M = 47, (SD=1.06)) outperformed on their CBT session (CBT/M=56, (SD=4.27)). However, the persons who preferred CBT did better than the other preference groups on PBT. But, to compare the prior testing mode and testing performance of On-Computer preference group, it was revealed that those test takers who preferred their exam in CBT version did not have a better performance on their CBT version. Also, from Table 6, by examining the relationship between prior testing mode preference and testing performance of On-Paper preference group, it can be seen that those test takers who preferred taking the test on PBT version (PBT/M=40.57, (SD=8.34)) (Table 5), outperformed on their CBT exam (CBT/M=41.42, (SD=15.95) (Table 6). The findings revealed that there was neither significant effect nor interaction between prior testing mode preference and their testing performance on either of the testing modes.

The feelings of test takers towards two versions of the same test and the impressions that they developed towards CBT after being exposed in the study were examined by distributing a researcher-made questionnaire. The analysis of the internal consistency resulted in an accepted reliability coefficient (N=120, Items=10, α=87). The responses of all 120 test takers to the statements of the questionnaire are displayed in Table 7.

As the table indicates, more test takers developed a positive attitude towards features of CBT. For example, it was easier to navigate through the PBT questions for 35% of the test takers while for 45% of the test takers; it did not vary to read the question in PBT or CBT. The greatest percentage for statement three was for the persons whose responses were no different. However, for 67.5% of the participants, it was easier to record their answers in CBT than in PBT while 67.5% found it easier to review their answers in PBT than in CBT.

Furthermore, 42.5% of the test takers found changing their answers easier in CBT than in PBT while 55% and around 33% found the CBT and PBT versions of the test more comfortable to take, respectively. From the table, it was concluded that more than 55% of the test takers guessed they would receive the same score on the CBT

version of the test. Interestingly, 65% of test takers enjoyed taking the test on CBT and more interestingly, 47.5% of the test takers thought that the CBT version of the test was more accurate to measure their vocabulary knowledge while only 10% of them responded that the PBT version could accurately measure their vocabulary ability. It is worth mentioning that these statistical analyses are compatible with the test takers' post-CBT preferences in testing group one and two. While 30% of the test takers preferred to take the PBT version of the test, 60% preferred taking the CBT version in the testing group.

Table 7. Attitudes of Participants towards both Testing Modes

N	questions	On paper in %		No difference in %		On computer in %	
		F.	***P.***	***F.***	***P.***	***F.***	***P.***
1	In which test, were questions and items navigated more easily?	15	35	51	42.5	54	32.5
2	In which test, were questions and items easier to read?	15	12.5	54	45	51	42.5
3	Which test was less fatiguing?	12	10	66	55	42	35
4	In which test, was it easier to record answers?	12	10	27	22.5	81	67.5
5	In which test, was it easier to review given answers?	39	67.5	-----	-----	81	32.5
6	In which test, was it easier to change answers?	39	32.5	30	25	51	42.5
7	Which test was more comfortable to take?	39	32.5	15	12.5	66	55
8	In which test, would you be more likely to receive the same score if you took it a second time?	12	9	42	35.5	66	55.5
9	Which test was more enjoyable to take?	42	35	-----	-----	78	65
10	Which test more accurately measured your vocabulary knowledge?	12	10	51	42.5	57	47.5

According to the post-CBT simple questionnaire responses of 60 participants of the testing group who were invited to have interviewed, 82.5% preferred computerized test and 17.5% showed a preference for paper-based test. In the interview, the participants confirmed their answers to the post-CBT and post-PPT testing mode preference questionnaire, i.e., would you prefer taking a test on paper/ no difference/ onscreen and then elaborate on their feelings and impressions of CBT and PBT and attitudes towards computerized counterpart of the conventional test. As the results of the quantitative part, the results of the post-survey analysis showed no correspondence between testing mode preference of test takers and their testing performance on the CBT version. The results of quantitative data revealed that those test takers of the testing group who preferred to take PBT (Table 5) outperformed in CBT (Table 6) and those who preferred CBT performed the same on both versions of the test. Based on the qualitative results, most of the participants showed high CBT preference as well as more advantages for CBT over PBT to rationalize why they prefer this mode of testing. 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 mode and the questionnaire on their attitudes towards the features of PBT and CBT. 100% of the participants who favored CBT mentioned "Easy to read items," "Easy to choose answers," Easy to change answers," and "Immediate scoring reports" as the advantages

to choosing CBT as their preferred testing mode. More than 78%, 60%, and 57% of the CBT advocators had positive attitudes towards the CBT features including "Enhanced security," Faster decision making as the result of immediate scoring and reporting," and "less time and effort" to take this format of the test, respectively.

Despite the high percentage of CBT preference 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 circling 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 one if they were on question 35, for example.

5 Conclusion

The received results and two sets of scores of test takers have been analyzed by the statistical package to find out any statistically significant difference between the two modes. Although several researchers have concluded that CBT version of the test resulted in lower scores than paper-based tests on participants' achievement (e.g. [34]), analysis of participants' testing performance in both PBT and CBT revealed that there was not any significant difference between the two sets of scores obtained from two formats of the test, and the test scores of participants were not different in paper-based and computer-based versions of the test. Test scores of participants did not vary in both PBT and CBT. Then the findings of the present research on score equivalence of two versions of the same test are in line with some studies that declare that two versions of the test are comparable (e.g., [27] [23] [35]). The findings are also in contrast to the findings of some others who claim that they are not comparable [29] [7] [36] [21].

Another main purpose of the study was to investigate the difference between the testing performance of male and female participants who took both PBT and CBT versions of the test. As the findings revealed, no significant difference was found for male and female participant groups' scores across the modes. The results of present research which included both male and female participants were compatible with the results that [37] reached. The findings of the current research on gender difference in testing mode comparability study were compatible with the findings of some research [22].

For analyzing research question two which focused on testing mode preference, Pearson Correlation, as well as descriptive statistics, were used. 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 testing performance. There was a weak negative correlation between both pre and post-CBT mode preference and CBT performance of the testing group. Moreover, the overall descriptive statistics of prior testing mode preference and testing performance of different preference groups' analysis answered negatively the research question two. These findings

indicated that there was not necessarily positive interaction between testing mode preference and testing performance. The reason might be either the testing orders, i.e., administration of CBT in the first testing session for a testing group or the novelty of CBT in the target setting [12].

6 References

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

[2] Alamri, M. (2019). Undergraduate students' perceptions of social media usage and academic performance: a study from Saudi Arabia. *International Journal of Emerging Technologies in Learning (iJET)*, *12*(10), pp. 35-55. https://doi.org/10.3991/ijet.v14i03.9340.

[3] Bennett, R. E. (1999). How the internet will help large-scale assessment reinvents itself. *Education Policy Analysis Archives*, 9(5), 1-25.

[4] Khoshsima, H. &HashemiToroujeni, S.M. (2017h). computer-based testing: score equivalence and testing administration mode preference in a comparative evaluation study. *International Journal of Emerging Technologies in Learning (iJET),14*(3), pp. 61-79. https://doi.org/10.3991/ijet.v12i10.6875.

[5] Chalhoub-Deville, M., & Deville, C. (1999). Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics*, 19, 273-299.

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

[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] Powers, D. E., & O'Neill, K. (1993). Inexperienced and anxious computer users: coping with a computer-administered test of academic skills. Educational Assessment, 1, 153–173.

[9] International Test Commission. (2006). International guidelines on computer-based and Internet-delivered testing. *International Journal of Testing*, 6, 143–171.

[10] HashemiToroujeni, 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 Art in Teaching*. Chabahar Marine and Maritime University (Iran).

[11] Johnson, M., & 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).Available from http://www.jtla.org.

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

[13] HashemiToroujeni, 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 Art in Teaching*. Chabahar Marine and Maritime University (Iran).

[14] Vansickle, T. (2015). Test Reliability Indicates More than Just Consistency. QUESTAR ASSESSMENT, INC.

[15] Bachman, L. F. & Palmer, A.S. (2000) *Language Testing in Practice*. (3rd ed.). Oxford University Press. The UK.

[16] Khoshsima, H. &HashemiToroujeni, 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). *24*(1), pp 80-99.
http://ijcjournal.org/index.php/InternationalJournalOfComputer/article/view/825/4188.

[17] Wells, S.C. &Wollack, J.A. (2003). An Instructor's Guide to Understanding Test Reliability. Testing & Evaluation Services publication, University of Wisconsin. Retrieved from http://www.wiscinfo.doit.wisc.edu/exams/Reliability.pdf.

[18] Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? Ergonomics Retrieved, *51*(9), 1352–1375 (p. 1362).
[19] Khoshsima, H. &HashemiToroujeni, 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), 1(2), February 2017; http://ijllalw.org/Current-Issue.html.
[20] Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), Educational Measurement (pp. 367–407). Washington, DC: American Council on Education.
[21] Khoshsima, H. &HashemiToroujeni, 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). http://dx.doi.org/10.5281/zenodo.268576.
[22] Khoshsima, H., Hosseini, M. &HashemiToroujeni, 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, 10(2) January (2017); http://dx.doi.org/10.5539/elt.v10n2p23.
[23] Khoshsima, H. &HashemiToroujeni, 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) *24*(1), pp 80-99. http://ijcjournal.org/index.php/InternationalJournalOfComputer/article/view/825/4188.
[24] Wells, S.C. &Wollack, J.A. (2003). An Instructor's Guide to Understanding Test Reliability. Testing & Evaluation Services publication, University of Wisconsin. Retrieved January 4, 2006, from http://www.wiscinfo.doit.wisc.edu/exams/Reliability.pdf.
[25] CBT McGraw-Hill (2003). Computer-based testing – Issues and considerations [Abstract]. It is retrieved from http://ccssocbtool.com/pdf_files/Design%20Considerations%20Article%201.pdf.
[26] 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.
[27] 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.
[28] Yurdabakan, I., &Uzunkavak, C. (2012). Primary school students' attitudes towards computer-based testing and assessment in turkey. Turkish Online Journal of Distance Education, 13(3), 177-188.
[29] Bennett, S., Maton, K. &Kervin, L. 2008. The "Digital Natives" Debate: A Critical Review of the Evidence. British Journal of Educational Technology, *39*(5), 775-786. https://doi.org/10.1111/j.1467-8535.2007.00793.x.
[30] Gallagher, A., Bridgeman, B., &Cahalan, C. (2000). The effect of computer-based tests on racial/ethnic, gender, and language groups (GRE Board Professional Report No. 96–21P). Princeton, NJ: Education Testing Service.
[31] Terzis, V., & Economides, A. A. (2011). The acceptance and use of computer-based assessment. Computers & Education, 56(4), 1032–1044. https://doi.org/10.1016/j.compedu.2010.11.017.
[32] Fowler, W.S. & Coe, N. (1976). Nelson English Language Tests. Canada: Thomas Nelson and Sons Ltd.
[33] Mon Mon The. (2018). The effectiveness of E-learning Experience through Online Quizzes: A Case Study of Myanmar Students. International Journal of Emerging Technologies in Learning (iJET), *13*(12), pp. 157-176.
[34] Mazzeo, J., & Harvey, A.L. (1988). The equivalence of scores from automated and conventional educational and psychological tests (College Board Report No. 88-8). New York: College Entrance Examination Board.
[35] 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. https://doi.org/10.1177/0013164406288166.
[36] Hosseini, M., ZainolAbidin, M. J., Baghdarnia, M., (2014). Comparability of Test Results of Computer-Based Tests (CBT) and Paper and Pencil Tests (PPT) among English Language Learners in Iran. International Conference on Current Trends in ELT, 659-667. https://doi.org/10.1016/j.sbspro.2014.03.465.
[37] Eid, G. K. (2004). 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.

7 Authors

Mohammad Reza Ebrahimi is working at Gonabad University, Gonabad, Iran.

Seyyed Morteza Hashemi Toroujeni has an M.A in TEFL. He works in the English Language Department, Faculty of Management and Humanities at Chabahar Maritime University, Chabahar, Iran. Email - hashemi.seyyedmorteza @ gmail.com

Vahide Shahbazi works at Chabahar Maritime University, Chabahar, Iran.

Article submitted 2019-01-18. Resubmitted 2019-02-26.Final acceptance 2019-02-27. Final version published as submitted by the authors.