Computer-Based Testing: Score Equivalence and Testing Administration Mode Preference in a Comparative Evaluation Study

https://doi.org/10.3991/ijet.v12.i10.6875

Hooshang Khoshsima, Seyyed Morteza Hashemi Toroujeni(✉)
Chabahar Maritime University, Chabahar, Iran
Hashemi.seyyedmorteza@gmail.com

Abstract—The empirical evidences show that two identical Computer-Based Testing (henceforth CBT) and Paper-and-Pencil-Based Testing (henceforth PBT) do not always result in the same scores. Such conclusions are referred to as "the effect of testing administration mode" or "testing mode effect". Moderators such as individual differences (e.g., prior computer experience or computer attitude) have been investigated [4] to see if they influence test takers' performance. The Guidelines for Computer-Based Tests and Interpretations [1] recommended eliminating the possible effects of some moderator variables on test takers performance. This research was conducted to provide the required empirical evidences on the existence of distinctive effects caused by changing administration mode from conventional PBT to modern CBT. The relationship between testing mode preference on test takers' CBT performance was also examined. Two equivalent tests and two questionnaires were used. Using descriptive statistics and ANOVA, the findings demonstrated that two CBT and PBT sets of scores were comparable. Additionally, prior testing mode preference and gender had no significant effect on test takers' CBT score, and they were not considered the variables that might affect the performance on CBT.

Keywords—Computer-Based Testing, Paper-and-Pencil-Based Testing, Testing Mode Effect, Gender Difference, Testing Mode preference

1 Introduction

As computers become increasingly available in educational contexts, it is likely that teachers will use them to administer tests. Using technological assessment tools to create tests has made it possible that tests be implemented and scored through computers. Tests that are created, implemented and scored via computers are called Computer-Based Testing (CBT). But, why CBT?. The constant development of the efficiency of available testing approaches that are achieved through enhancement of computers' effectiveness to record, gather and analyze test scores caused improvement and extension of computerized testing [2]. Increasing use of computer as a mass-consumed commodity in early 1990s [3] caused acceleration of development of computing technology. In fact, such developments which leaded to the constant use of

CBT in educational contexts due to its great efficiency and practicality contributed to stabilize its predominant role in assessment field [4], [5]. Computerized test can be manipulated in such a way that gives the test takers the possibility to choose when they prefer to participate in a testing session [6]. CBT saves time of supervising and marking [7]. CBT provides more flexible and comfortable testing environment than PBT, and the accuracy of computer in scoring and reporting the results is better due to more effective functions of personal computers [8]. Detailed automatic feedback can be given to test takers as soon as the test is terminated [6], [9], [10], [11]. Although manipulating and marking computerized testing are done easily [11], this kind of testing is not fundamentally and intrinsically better than its PBT counterpart [12]. Before developing CBT version of a test or converting the PBT version into its CBT counterpart, the critical issues of reliability and validity have to be measured.

In contrast to the traditional form of tests in which organizational limitations are imposed on tests administered in groups, CBT kind of tests can be individually administered [13], [14]. This is especially true for a subcategory of CALT called computer adaptive language test. According to Chapelle and Douglas, validity of CBT counterpart of PBT version may be violated such that its results might not match the results obtained from PBT. It means that the similarity of scores and results received from two versions of the same test guarantee the validity of the tests. On the other hand, another way based on which the validity issue may be violated is that the created question items presented onscreen may be different from those created and presented in other formats. Thereby, testing practitioners should notice that the conceptions related to the validity of computerized testing may not be necessarily related to the validity of other formats of the test such as PBT version. In fact, the applicability of the validity in different versions of tests may be different due to the unique features of each format [15].

The main purpose of the present research was to study the comparability of the test scores received from CBT and PBT versions of an equivalent test. Furthermore, the relationship between variables that may affect the students' performance on tests was also examined. As it has been examined by other researchers and mentioned in the literature, the researcher of the present study needs to check the consistency of the findings with previous ones to support or reject the related hypotheses. The researcher actually tends to check whether or not test takers prefer or do not prefer the computerized test, and how this testing mode preference interact their performance on CBT. The researcher hopes the present study expect that this piece of research work succeed in adding to the knowledge in comparability study in academic contexts.

2 Literature review

Applying computerized versions of test in educational contexts is increasing [16], [17], [18], [19], [20]. Guidelines for Computer-Based Testing were published by Association of Test Publishers (2000) [21], International Test Commission (ITC) (2004) [22] (International Guidelines on Computer-Based and Internet-Delivered Testing), American Council on Education (1995) [23] (Guidelines for Computerized

Adaptive Test Development and Use in Education). Among the guidelines published by several organizations, the International Test Commission (2004) [22] and the American Psychological Association (1986) [1] devoted their standards and guidelines to CBT exclusively. The International Test Commission claims that test designers should establish the comparability between internet-based testing and its traditional counterpart [24]. To establish and produce relatively equivalent reliability and correlation, means, and standard deviations, as well as the ways to gather required data from the same subgroups of examinees in order to achieve the comparability and equivalency based on the same versions of the test are the major goals of the published guidelines.

To investigate the equivalence of scores obtained from PBT and CBT, many studies have been done in language testing area. Some of these studies examined the score equivalence of computerized version of a reading test and its conventional counterpart. One of these studies is the research done by Mark Pomplun, Sharon Frey, and Douglas F. Becker in 2002. Pomplun, Frey & Becker (2002) [25] studied the score equivalence of currently used paper-and-pencil version of a Nelson-Denny speeded test of reading comprehension and a new computerized version. Their findings showed that the reading onscreen was more difficult than PBT but it takes shorter time to complete. They declared that difficulties in onscreen reading were due to the issues of primitive technology. Furthermore, they concluded that the scores received from CBT were higher than the scores obtained from PBT [25]. According to Bugbee (1996) [26], two CBT and PBT versions of a test can be equivalent if one of the following criteria is met. The criteria that should be met to name two versions of a test equivalent are (a) equal means and distributions of two versions (b) equal means and distributions, reliabilities and correlations with criterion (validity) variables for interchangeable test scores. Khoshsima & Hashemi Toroujeni (2017b) declare that computer-based testing is becoming popular as a green computing strategy due to its advantages over PBT. They conducted a research on comparability of computerized testing and paper-based testing. They also examined the relationship of some external moderator variables such as testing mode effect, testing mode order, computer attitudes and testing mode preference. Findings of the study indicated the interchangeability of scores of test takers from CBT and PBT. No statistically significant difference was found across modes. Moreover, the moderator variables whose relationship with CBT performance was examined were not considered external factors that might affect test takers' performance on CBT [27]. Lightstone and Smith (2009) [28] conducted a study on the comparison between PBT and CBT considering test mode preference and test performance on CBT. They collected and analyzed their data through two studies including asking the students to explain which test they preferred to take more and asking them to explain their reasons for their choices and their prediction of their test scores according to their preference. Like Khoshsima, Hosseini and Hashemi Toroujeni (2017), the researchers investigated the effect of testing mode preference on test results. The results of these two related studies showed that the students' justification for choosing the test format differ among individuals and both test mode preference and test performance are dependent to some extent on individual differences and meta-cognitive factors such as cognitive workload in test taking. They showed high

preference for computerized tests; however, their prediction on getting better score on CBT was not true as they performed better on PBT [5]. Khoshsima, Hosseini and Hashemi Toroujeni (2017) investigated the correlation of testing mode preference with CBT performance. Their findings revealed no correlation between testing mode preference and testing performance of test takers on computerized version of the test [5]. Their findings were compatible with the findings of the previous research done by Flowers et al. (2011) in which there was a high preference for CBT, but test takers' preference had negative correlation with their performance on CBT [29]. Similar to the aforementioned studies, no significant difference was found between test takers' scores on two versions of test which indicated no correlation between testing mode preference and test performance [30].

The present research examined the relationship of testing mode preference with CBT performance. Some researchers have done similar studies on the student's preference on delivery testing mode [31], [32], [33], [20]. The participants of all these studies were asked through either interviews or questionnaires which tests they preferred more: computerized or paper-based one. Moreover, some of these studies investigated whether test takers' testing preference would influence their testing performance on CBT or not. The results of most of these studies demonstrated the preference for computerized tests. Yourdabakan and Uzunkavak (2012) investigated the attitudes of 784 students from both private and state schools in Turkey and found that the students from public schools were more favorable of CBT than students from privates schools; however, test results were not different, which implied no correlation between students' test preference and test results on CBT [20]. In parallel with Al-Amri (2009), Escudier et al. (2011), Yourdabakan and Uzunkavak (2012), and Hashemi Toroujeni (2016) also found the high preference for computerized tests among their participants with no difference in test scores of two test types. They also confirmed no correlation between test scores and testing mode preference [31], [32], [20], [4].

To examine the relationship between test takers' preference and their test scores, Hashemi Toroujeni (2016) and Khoshsima, Hosseini & Hashemi Toroujeni, (2017) utilized either preference scale questionnaire or interviews to ask which testing mode of administration they prefer [4], [5]. The researcher of the current study, like some of the aforementioned researchers, investigates the influence of test takers' preference on their testing performance on CBT in details. In a study, Higgins et al. (2005) examined which test 161 participants preferred to take. The findings showed that 87% of participants preferred to take CBT due to ease of use feature of CBT kind of testing. No significant difference was also found between test takers' scores on two versions of a test which indicated no correlation between test mode preference and test performance [30]. In another study done by Al-Amri (2009), although test takers preferred to take CBT, their test performance was better on PBT. His research findings show no relationship between test performance and testing mode preference [31]. Another research studying the preference of testing mode indicated that children performed better on PBT than CBT but they preferred to use computer as the medium through which they prefer to take their tests [34].

The effect of individual differences and backgrounds on test takers' performance is a critical issue regarding validity of CBT [4], [12], [27], [35], [11]. Leeson (2006) defines some variables including user variables and technology variables that are the major sources of difficulties in implementing CBT [36]. User's gender, capability of processing information, the ability to use a computer, prior attitudes, anxiety and etc. may have definite impacts on CBT performance of test takers. In some studies, investigating the difference between methods in terms of gender, race and age [37], [38] leaded to no significant difference in achievements; while in other studies [39] little significant difference was observed. In their recent study, Terzis and Economides (2011) describe the trends of male and female students towards CBT [40]. In their study, Terzis and Economides (2011) and Khoshsima, Hosseini & Hashemi Toroujeni (2017) describe the trends of male and female students towards CBT [40], [5]. Khoshsima, Hosseini & Hashemi Toroujeni (2017) report no significant difference for male and female test takers' scores across the modes [5].

Among the objectives defined for the current research the principle purpose that has to be followed is to investigate whether there is any correlation between testing mode preference and testing performance. Before investigating this relationship, the researcher examined score equivalence of testing groups. Then, in order to achieve the aforementioned objectives, the following research questions have been addressed.

1. Is there any statistically significant difference between computer-based language testing and paper and pencil-based testing?
2. Do participants' prior testing mode preferences affect their performance on CBT?
   (a) Do participants perform better on their preferred test mode?
   (b) Do participants' exposures to CBT influence their posterior testing mode preference?
   (c) Do CBT experiences result in a positive attitude towards features of CBT?

3 Methodology

3.1 Research design

Both quantitative and qualitative data was collected based on a mixed-methods approach including multiple choice achievement tests, questionnaires and interviews that were used in the present research as the methodological approach. Two versions of two equivalent tests were administered to the homogenous test takers who were assigned to two testing groups based on the *common person design* which is an efficient and practical design in detecting differences especially in smaller sample of test takers to collect good data for making score comparison. In this study, testing mode of administration was considered the treatment. Actually, to avoid any individual characteristics' influence on testing performance, within-subject group was also used in the study, i.e. the same group was given two equivalent tests in the form of CBT and PBT.

3.2 Participants

After administering PBT TOEFL General Proficiency Test to 116 graduate students of Chabahar Marine and Maritime University (CMU) to make sure about their proficiency level as upper-intermediate, 96 homogenous students were determined. Due to the lack of space in testing environment and the number of computers, 80 homogenous graduate students including 40 girls and 40 boys were randomly selected and assigned to two major testing groups. To get better results on the relationship of gender difference and testing performance, all the 40 girls were assigned to testing group 1 and all the 40 boys were organized in testing group 2. In addition to the two main testing groups that were supposed to participate in the main investigation and tests, those 16 homogenous students was used as the pilot group. This group that was similar to the other two testing groups was employed to do our pilot study and to examine the reliability of data collection instruments used in this research. From the 80 upper-intermediate graduate students as the test takers of the research who took the placement test in autumn of 2016, there were equal numbers of boys (n=40=50%) and girls (n=40=50%). The age range of all the participants was between 22 to 35 years. Mean age was 25.5 years with a standard deviation of 4.63. All the boys were organized in one testing group, and all the girls were organized another one to compare the mean score of male and female to examine the gender difference on CBT.

3.3 Instruments

Since the quality of any research depends heavily on the quality of gathered data and the procedures to gather data, it is essential to conduct the research in the most appropriate and suitable framework [4], [41]. The researcher used two groups in a counter-balanced order to investigate the effect of testing mode and gender difference on test takers' performance. Two groups took two computer-based and paper-based versions of two equivalent tests. In paper-based test, test takers should read each question and mark the right option on the separate answer sheet that was given to each test taker with the test papers.

The web-based CBT version of the test with fifty questions examining test takers' vocabulary knowledge for the educational measurement was developed. Test takers were required to read a question appearing on the computer screen and choose the most appropriate option under each question by clicking the mouse on the blank space besides the options. The next research data collection instruments were a simple question appeared at the bottom of exam paper and screen, i.e. *would you prefer taking test on paper – no difference – computer* to examine the relationship between testing mode preference and performance as well as a researcher-made questionnaire entitled Attitudes towards the Features of Computer and Paper-based Testing (AFCPT). As the last instrument, interviews were done to collect the qualitative data. Furthermore, SPSS version 22 software running on a CORE i5 Sony VAIO laptop was utilized to compare the obtained results from two formats of the test as pretest and posttest in the experiment.

3.4 Procedure

The first section of the research was designed to examine the effects of testing mode scores obtained from CBT as their performance on computerized version of the test and to compare the received results with its PBT counterpart's results. Two equivalent multiple-choice tests have been prepared from the course book of General English Book. Each exam consisted of 40 multiple-choice items. The two selected test sets have been piloted to ensure the equivalency of two tests in relation to reliability and validity. It means that the two selected tests have tried out with an adequate number of test takers who possessed similar characteristics as the target population to ensure the equivalency of tests according to the degree of reliability, mean scores, and standard deviations.

The second study which employed both quantitative and qualitative instruments including questionnaires to collect research data was designed to investigate the possible effects of testing mode preferences on test takers' performance. The interview as the qualitative instrument was used to gather more research data on the features of two versions of the test that the test takers have taken. So, when testing group one was taking the PBT form of the test in one of the classes of Language Center of CMU, testing group two was taking CBT form of the same test in Information Technology Center of CMU. After four weeks interval, two testing groups two versions of the test in a reversed order. According to Kauffman (2003), although no minimum or maximum interval between two test administrations was suggested, four weeks interval was determined between implementing two versions in this research [42]. A four weeks interval could be an adequate minimum period to wait before conducting the second testing session. Kauffman (2003) declares that both short and long periods might affect testing performance by resulting in practice effect and maturation or instruction, respectively [42].

Testing group one and two took the CBT and PBT forms of the tests, respectively. In each CBT testing session, before implementing the CBT version of the test, a simple short researcher made test as well as some oral instructions on how to answer this type of the test were given to the test takers.

To investigate if any change was occurred in test takers' preference toward taking paper-based or onscreen test, test takers were asked to answer a simple question mentioned at the bottom of their exam paper and screen, i.e. *would you prefer taking test on paper – no difference – computer* to examine the relationship between testing mode preference and performance. This question studied the possible change in test takers' preference after taking CBT version of the exam. Furthermore,

In addition to the simple question mentioned at the end of both versions of the exam, the feelings and impressions of test takers about CBT mode of administration were studied after their exposure to CBT by another researcher-made simple questionnaire (AFCPT) to collect the required data to analyze research questions 2, 2.1, 2.2, 3.3. This instrument that was a set of researcher-made questions regarding the testing mode preference assessed the development of positive or even negative attitudes towards CBT and collected some parts of remaining research data.

In the last stage of the research and after filling out the questionnaires and to confirm the data obtained from questionnaires, 20 participants who filled out the questionnaires were randomly selected from among the volunteers for interview. They were encouraged to expand on their answers to the questionnaires by offering some reasons for their reported mode preferences. Bachman and Palmer advised that reliance on only questionnaires may restrict test takers' responses so that researchers only get answers to the questions they ask. In contrast, open-ended interview questions may produce unexpected responses from test takers which would hopefully lead to new insights into the phenomenon under investigation [43].

4 Results and discussion

According to Warner (2013), one of the main purposes of doing pilot study in which the research data collection instruments are tried out is to get the correct feedbacks of conducting the main test and overcome the possible unexpected problems during the procedure [44]. In the present research, those who participated in the pilot study were the selected homogenous participants who had not the opportunity to take part in the main investigation. Creswel & Clark (2007) asserts that the participants of the pilot study should be as similar as possible to the main participants [45].

Paper-based versions of the two prepared multiple-choice tests were administered to the participants of the pilot study in two separate testing sessions with the interval of two weeks during the educational semester to ensure the reliability of the items. In addition to comparing the mean and standard deviation of both tests administered to pilot group in two testing sessions, Cronbach's Alpha Coefficient was also employed to calculate the reliability. The results of determining the equivalency of two tests revealed a high correlation (Table 1). Then it was concluded that both tests were similar enough to be considered equivalent [1].

Table 1. Statistical tendency of pilot study

Multiple choice-tests	Number of test takers	Mean	Std. Deviation	Cronbach's Alpha Coefficient
Test 1	16	25.32	4.3	.906
Test 2	16	24.97	4.8	.913

Since parametric statistical tests are based on some assumptions, the researcher had to confirm fulfillment of four assumptions of interval data, independence of subjects, normality distribution and homogeneity of variances. The collected data i.e. exam scores were measured on an interval scale. The testing performance of 60 participants who were assigned to two testing groups to take two versions of the equivalent tests in four testing sessions was independent of each other and no treatment by peer or group work was administered in this research. Furthermore, since normal data is the fundamental assumption in parametric statistical testing [46], the present study checked the assumption of normality as well as homogeneity of variances.

Table 2 displays the results obtained from two statistical tests of normality namely Shapiro-Wilks and Kolmogorov-Smirnov. Wilk test is the most powerful test to esti-

mate normal distribution of small sample sizes [47]. From Table 3, p-values were greater than 0.05. Then, as the Sig. values under Shapiro-Wilk were >0.05, (.564, .531, .795, and .801 for TG1 PBT, TG1 CBT, TG2 PBT and TG2 CBT, respectively), we concluded that the related variables were normally distributed.

Furthermore, Levene's Test of Homogeneity of Variances was run; the result, F (3, 117) = 7.7, p = .716, with an alpha level of .05, p (.716) 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 (.716) >α (.05). It means that our data had similar variances and we need using parametric statistical tests.

For the first analysis, the one-way analysis of variance (ANOVA) was used to compare the means of two sets of scores of each group (related group) obtained in two different testing sessions. To gain a better view of the data, descriptive statistics and then inferential statistics analysis were used to find out the relationship between mean scores. According to the results, testing group one mean score on CBT (M = 46.13, SD = 13.80) was higher than that group's mean score on the PBT (M = 46.66, SD = 17.43). Testing group two mean score on PBT (M = 45.40, SD = 21.2) was higher than that group's mean score on CBT (M = 39.6, SD = 13.15). Additionally, of the two CBT sessions of the tests taken by testing, the highest mean score was found in CBT session of testing group one; with a relatively higher mean score by 7 points (Table 3).

To answer the main question of the research that looks for the significance difference between the scores of two versions of the tests, One-Way analysis of variance

Table 2. Testing Normality Assumption

	Tests of Normality					
	Kolmogorov-Smirnova			**Shapiro-Wilk**		
	Statistic	***D.F.***	***Sig.***	***Statistic***	***D.F.***	***Sig.***
TG1 PBT	.117	33	.871	.943	33	**.564**
TG1 CBT	.185	33	.795	.901	33	**.531**
TG2 PBT	.305	27	.803	.747	27	**.795**
TG2 CBT	.318	27	.867	.855	27	**.801**

(TG=Testing Group)

Table 3. Distribution of participants' scores in PBT & CBT

Descriptive statistics								
	N	**Mean**	**Std. Deviation**	**Std. Error**	**95% Confidence Interval for Mean**		**Minimum**	**Maximum**
					Lower Bound	***Upper Bound***		
TG1 PBT	33	**46.66**	**17.43**	3.18	40.15	53.17	26.00	98.00
TG1 CBT	33	**46.13**	**13.80**	2.52	40.97	51.28	20.00	64.00
TG2 PBT	27	**45.40**	**21.20**	3.352	38.61	52.18	26.00	98.00
TG2 CBT	27	**39.60**	**13.15**	2.07	35.39	43.80	20.00	64.00

was conducted with a null hypothesis of no difference. According to the results of One-Way ANOVA test (Table 4), 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 four testing sessions, the Sig. value was .896 at P<0.05. This amount of significance value at 117 (N-3) 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).

According to the descriptive statistics, female test takers' mean score on CBT (46.13) was higher than the mean score of testing group two on CBT (M = 39.60, SD = 13.15). On the other hand, the standard deviation in testing group one CBT was a bit higher than the CBT performance of testing group two. It means that the dispersion of scores from mean score in CBT of testing group one was higher than in CBT of testing group two; consequently, it was concluded that Standard Error of Measurement (SEM) in testing group two CBT was lower than in testing group two CBT (Table 5).

Table 4. One-Way ANOVA comparing scores of participants in PBT & CBT

ANOVA					
	Sum of Squares	***D.F.***	***Mean Square***	***F***	***Sig.***
Between Groups	4.267	3	4.267	.017	**.896**
Within Groups	14338.133	118	247.209		
Total	14342.400	119			

Table 5. Distribution of CBT scores of two testing groups

Descriptive Statistics								
	N	**Mean**	**Std. Deviation**	**Std. Error**	**95% Confidence Interval for Mean**		**Minimum**	**Maximum**
					Lower Bound	***Upper Bound***		
CBT1	33	**46.1**	**13.80**	2.5	40.97	51.28	20.00	64.00
CBT2	27	**39.6**	**13.15**	2.0	35.39	43.80	20.00	64.00

Of the male (Testing Group 2) and female (Testing Group 1) CBT sessions of the tests, the highest mean score was found in female CBT, with a relatively higher mean score by 1 point. According to the descriptive statistics for two CBT results, it was revealed that the female test takers outperformed on computerized testing. In fact, there was a difference between mean score of male and female testing groups, and the difference, according to posteriori test, was not statistically significant.

Tamhane's T2 post hoc test displayed that there were no statistically significant differences between mean score of PBT and CBT versions of testing group one, mean score of PBT and CBT versions of testing group two, and between mean score of CBT version of testing group one and CBT version of testing group two that were indicated by (p = .872), (p = 524), and (p = .124), respectively. Additionally, *Games-*

Howell post hoc statistical test for examining variables of the groups revealed that there were no statistically significant differences between PBT and CBT versions of testing group one, PBT and CBT versions of testing group two, and between CBT version of testing group one and CBT version of testing group two that were indicated by (p = .959), (p = .375), and (p = .177), respectively.

The relationship of pre- and post-CBT mode preference of the test takers of the first testing group with their testing performance on CBT version of the test was measured by Pearson's product-moment correlation statistical test. There was a weak negative correlation between both pre and post-CBT mode preference and CBT performance of testing group one, r (28) =.-.016, $p < .916$ and r (28) =.-.121, $p < .472$, respectively. Furthermore, the Pearson's product-moment correlation to assess the relationship of post-CBT and post-PBT mode preference with CBT performance of all the test takers of testing group two revealed that there was a weak negative correlation between both post-CBT and post-PBT mode preference and CBT performance of testing group two, r (28) =.-.083, $p < .452$ and r (28) =.-.113, $p < .386$, respectively (Table 6).

In the next step, descriptive statistics of different testing mode preference groups were used to gain a better view of the data and to find out the relationship between mean scores of testing groups. The descriptive statistics output is displayed in Tables 7, 8, and 9.

According to Table 7, the mean score obtained for the test takers who preferred On-Computer test (these test takers organized in a group called On-Computer preference group) (PBT1 / M = 64, (SD = .0)) was absolutely greater than the mean scores received for all the other test takers who were organized into two other groups including No-Difference and On-Paper preference groups. Based on the results, those test takers who preferred taking CBT over its paper-based counterpart had better performance (PBT1 / M = 64, (SD = .0)) on paper-based test compared to those whose preference was PBT version (PBT1 / M = 40.57, (SD = 8.34)).

Additionally, the CBT and PBT results of various testing mode preference groups (between groups) are indicated in Table 7. Based on the acquired results, the test takers whose preference to take the test on either CBT or PBT versions was taking the test on its PBT format (PBT1 / M = 40.57, (SD = 8.34)) had better performance on CBT version compared to PBT one (CBT1 / M = 41.42, (SD = 15.95)) (Table 8). Then, the PBT mean score of the test takers of the first testing group who preferred CBT counterpart (PBT1 / M = 64, (SD = 0)) (see Table 7), did not vary with their CBT mean score (CBT1 / M = 64, (SD = 0)) (see Table 8). Accordingly, the test takers of No-Difference testing mode preference group, outperformed on their CBT session (PBT1 / M = 47, (SD = 1.06)) (Table 7), (CBT1 / M = 56, (SD = 4.27)) (Table 8). However, the overall results of prior testing mode preference and testing performance of different preference groups' analysis answered negatively the research question 2. These findings indicated that there was no 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 [48], [5].

Table 6. Pearson Correlation of pre-CBT and post-CBT mode preference with CBT scores of testing group one and post-CBT and post-PBT mode preference with CBT scores of testing group two

Pearson Correlations		***Pre-CBT Mode Preference***	***Post-CBT Mode Preference***	***Post-CBT Mode Preference***	***Post-PBT Mode Preference***
CBT1	Pearson Correlation	**-.016**	**-.121**		
	Sig. (2-tailed)	**.916**	**.472**		
	N	30	30		
CBT2	Pearson Correlation			**-.083**	**-.113**
	Sig. (2-tailed)			**.452**	**.386**
	N			30	30

Table 7. PBT performance of different preference groups of testing group one

Pre-CBT Mode Preference	**N**	**PBT 1 Mean Score**	**Std. Deviation**	**Std. Error**	**95% Confidence Interval for Mean**		**Minimum**	**Maximum**
					Lower Bound	***Upper Bound***		
On Paper	28	40.57	8.34	1.57	37.3368	43.8061	28.00	52.00
No Difference	8	47	1.06	.377	46.1063	47.8937	46.00	48.00
On Computer	4	64	.00	.000	64.0000	64.0000	64.00	64.00
Total	40	44.20	9.98	1.57	41.0074	47.3926	28.00	64.00

Table 8. CBT performance of different preference groups of testing group one

Pre-CBT Mode Preference	**N**	**CBT 1 Mean Score**	**Std. Deviation**	**Std. Error**	**95% Confidence Interval for Mean**		**Minimum**	**Maximum**
					Lower Bound	***Upper Bound***		
On Paper	28	41.42	15.95	3.01	35.2429	47.6142	14.00	66.00
No Difference	8	56	4.27	1.51	52.4250	59.5750	52.00	60.00
On Computer	4	64	.00	.00	64.0000	64.0000	64.00	64.00
Total	40	46.6	15.74	2.48	41.5652	51.6348	14.00	66.00

Then, the researcher examined how was the performance of different preference groups of testing group one on CBT version of the test (the second version that the group took in its second testing session). As it was shown in the Table 8, in the second testing session of testing group one i.e. CBT version, the CBT mean score of On-Computer preference group (CBT1 / 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 (CBT1 / M = 41.42, (SD = 15.95)) on CBT version of the test. On the other hand, those who didn't mind taking the test on either

modes (PBT1 / M = 47, (SD = 1.06)) outperformed on their CBT session (CBT1 / 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. And, from the Table 8, by examining the relationship between prior testing mode preference and testing performance of On-Paper preference group, it was concluded that those test takers who preferred taking the test on PBT version (PBT1 / M = 40.57, (SD = 8.34)) (Table 7), outperformed on their CBT exam (CBT1 / M = 41.42, (SD = 15.95) (Table 8) in testing group one. However, the findings indicated that there was no necessarily positive interaction between testing mode preference and testing performance. 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.

Table 9. CBT and PBT performance of different preference groups of testing group two

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	*Upper Bound*		
CBT2	No Difference	6	32	.00	.00	32.0000	32.0000	32.00	32.00
	On Computer	34	40.94	13.85	2.37	36.1057	45.7766	20.00	64.00
	Total	40	39.6	13.15	2.07	35.3939	43.8061	20.00	64.00
PBT2	No Difference	6	33.33	1.03	.42	32.2495	34.4172	32.00	34.00
	On Computer	34	47.52	22.36	3.83	39.7263	55.3325	26.00	98.00
	Total	40	45.4	21.20	3.35	38.6181	52.1819	26.00	98.00

According to Table 9, the changes of performance of different preference groups of testing group two on CBT version of the test (the first version that the male group took in its first testing session) and on PBT can be traced. As the Table 9 revealed, after implementing the first testing session of testing group two i.e. CBT version of the test, those participants who preferred taking CBT version of the test (CBT2 / M = 40.94, (SD = 13.85)) (Table 9) outperformed in their PBT exam (PBT2 / M = 47.52, (SD = 22.36)) (Table 9). And those who didn't mind taking the test on either mode (CBT2 / M = 32, (SD = 0)), did better on their PBT (PBT2 / M = 33.33, (SD = 1.3)).

According to the results, it was concluded that no one in testing group two preferred to take PBT after implementing CBT version of the test in testing session one. But it seems that some of them changed their opinions after taking PBT version of the test in the second testing session. This issue will be studied in the next paragraphs of this section.

To answer the research questions 2.2 and 2.3, testing mode preference of test takers of testing group one and two were examined before and after exposure to CBT. To achieve this goal, we asked the test takers of two testing groups to answer the simple preference mode questions before and after both versions of the test to investigate the

impact of exposure to CBT on their testing mode preference. Then, the testing mode preference was categorized under two pre-CBT and Post-CBT testing mode preferences. Additionally, although the CBT version of the test was administered to the test takers of testing group two in their first testing session, the question was asked to the test takers of this group to see if its test takers' testing mode preferences have been changed after taking the PBT version of the test. Thus, to measure each pre and post-CBT testing mode preference, the answers of the test takers to the first and second questionnaires were investigated. Table 10 indicates the frequency of test taker's responses before exposure to CBT.

According to the results, after implementing PBT version of the test and before the second testing administration session, 70% of the test takers of testing group one preferred to take the test on paper, and 20% of the test takers didn't mind taking the test in either mode. From Table 10, it can be seen that just 10% opted for computers as their preferred mode of testing.

To examine the reasons of their choosing, some of the participants were selected randomly to be interviewed after the testing sessions. To examine the possible change of test takers' testing mode preference, their responses to the second simple questionnaire administered after the CBT version of the test were examined too.

To see if any change has happened to the mode preference of test takers of group one, their answers to the second simple questionnaire were examined. As you can see in the Table 11, only 30% of the test takers still preferred PBT version of the test while just 10% didn't mind taking the test on either mode. The greater percentage 60% was the test takers who opted for computer as their preferred mode of testing. According to the results of Tables 11 and 12, we concluded that the number of participants who preferred PBT and who didn't mind taking the test in either mode have changed in favor of the test takers who chose On-Computer as their preferred testing mode preference.

In testing group one, it was concluded that the percentage of various preferred testing modes in the first testing session i.e. PBT version of the test was changed in favor of CBT after implementing this version of the test. Then, we examined the testing mode preferences of testing group two for which CBT was the first implemented test. To get a better view of the preferences of different preference groups in testing group one, we used the results shown in Table 15 and then compare the results with the responses of test takers to the simple questionnaire which was administered after PBT version in the second testing session.

According to the results, after implementing CBT version of the test in the first testing session of group two, interestingly, no one preferred to take the test on paper, while 15% of the test takers didn't mind taking the test in either mode. But, the greater percentage 85% was the test takers who preferred CBT as their preferred mode of testing.

After administering the simple questionnaire to the test takers in the second testing session, 15% of the test takers preferred PBT version of the test while just 15% didn't mind taking the test on either mode. Although the number of test takers who preferred CBT has decreased in 10%, the greater percentage 70% was still the test takers who opted for computer as their preferred mode of testing. 'According to the results of

Table 10. Frequency Table of responses to the Pre-CBT testing mode preference of testing group one

Pre-CBT1 Testing Mode Preference					
		Frequency	***Percent***	***Valid Percent***	***Cumulative Percent***
Valid	On Paper	28	70.0	70.0	70.0
	No difference	8	20.0	20.0	90.0
	On Computer	4	10.0	10.0	100.0
	Total	40	100.0	100.0	

Table 11. Frequency Table of responses to the Post-CBT testing mode preference of testing group one

Post-CBT1 Testing Mode Preference					
		Frequency	***Percent***	***Valid Percent***	***Cumulative Percent***
Valid	On Paper	12	30.0	30.0	30.0
	No Difference	4	10.0	10.0	40.0
	On Computer	24	60.0	60.0	100.0
	Total	40	100.0	100.0	

Table 12. Frequency Table of responses to the Post-CBT testing mode preference of testing group two

Post-CBT2 Testing Mode Preference					
		Frequency	***Percent***	***Valid Percent***	***Cumulative Percent***
Valid	No Difference	6	15.0	15.0	15.0
	On Computer	34	85.0	85.0	100.0
	Total	40	100.0	100.0	

Table 13. Frequency Table of responses to the Post-PBT testing mode preference of testing group two

Post-PBT2 Testing Mode Preference					
		Frequency	***Percent***	***Valid Percent***	***Cumulative Percent***
Valid	On Paper	6	15.0	15.0	15.0
	No Difference	6	15.0	15.0	30.0
	On Computer	28	70.0	70.0	100.0
	Total	40	100.0	100.0	

Tables 12 and 13, we concluded that the number of participants who preferred CBT has changed in favor of the test takers who chose On-Paper as their preferred testing mode preference, but most of the test takers still preferred taking the test on CBT. However, the interviews that we conducted after testing sessions gave the opportunity to the test takers to elaborate on their reasons to prefer a version of the test or the reasons of changing their testing mode preference choose. As well, the interviews provided us with good chance to identify the attributes of test takers' possible alteration from one testing mode to the other one.

We also examined the feelings of test takers about two versions of the same test and the impressions that they developed about CBT after being exposed in our study. To achieve this goal, the AFCPT questionnaire was administered to the test takers at the end of their second exam in the second testing session. According to the responses of all 80 test takers to the statements of the questionnaire, more test takers developed a positive attitude towards the 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 didn't vary to read the question in PBT or CBT. The greatest percentage for the statement three i.e. which test was less fatiguing?, was for the persons whose responses were no difference. 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. Based on the results, 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 PBT version of the test, 60% preferred taking the CBT version in testing group one. Moreover, in testing group two, 85% preferred CBT while just 15% liked the conventional format.

The qualitative research data that was collected to support the quantitative research data came from conducting a semi-structured interview with 20 participants who were randomly selected from two testing groups. 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. Once transcription of the data has been completed, content analysis was conducted on transcribed data by identifying all the main concepts. In thematic analysis, similar statements and responses to the same questions were coded and categorized under a common theme [49]. The 16 people who advocated the CBT and 4 ones who preferred PBT were asked some questions to rationalize their testing mode preference and explain their reasons to find out the rationales behind each preference to support the findings of quantitative analysis in attitudes.

Based on the results, most of the participants showed high CBT preference as well as more advantages for CBT over PBT to rationalize why they prefer this mode of testing. It could 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 AFCPT questionnaire on their attitudes towards the features of PBT and CBT. the most frequently cited advantages of CBT were (A) Faster test taking, (B) Fewer response entry and recognition errors, (C) Faster and more controlled test revision process with shorter response time, (D) Instant score report, (E) Enhanced security.

5 Conclusion

ANOVA statistical test was run to compare the means of two sets of scores of each group (related group) obtained in two different testing sessions. Based on the findings, it was concluded that there was no statistically significant difference in the mean scores of two testing groups in four testing sessions as a whole (p=.896). The findings of the research question one were compatible with the results of (Al-Amri, 2008; Khoshsima, Hosseini & Hashemi Toroujeni, 2017) [50], [5] who claim that assessments are comparable across modes. The findings were also in contrast to the other researchers (Hosseini et al. 2014) [41] who disagree with the comparability of scores obtained from two testing modes. By considering comparability studies in Iranian educational contexts, unlike Hosseini et al. (2014) [41], the findings of this study are in line with the findings of Khoshsima, Hosseini & Hashemi Toroujeni that supports the equivalency between test scores of PBT and CBT [5]. According to the results of the current piece of research work, the equivalency of scores received from two versions of the test in higher educational setting is confirmed.

Secondly, gender difference and testing mode preference had no significant relationship with CBT performance among Iranian graduate students studying in state university. In fact, it was demonstrated that the factors had not impact or interaction on computerized counterpart of PBT and there was no 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 [48], [5]. The findings of the present study were in consistent with the result of Khoshsima et al.'s (2017) [5] study that found out test takers with positive attitudes towards the use of computer did not perform better on CBT. Then, according to the findings of the present study supports that testing mode preference as an external moderator variable cannot be considered a factor that may have effect on the CBT performance. In fact, the possibility of existence of any relationship between testing mode preference and testing performance was rejected based on the findings. Moreover, the analysis of the qualitative data gathered in the interview sessions supports the findings of quantitative section of the research. In other words, although some students as the test takers of the main investigation demonstrated high testing mode preference towards the CBT version of the test, they didn't necessarily outperformed on the testing occasion in which the CBT version was implemented. This study added to the current research that focused on comparability issue between paper-based and computer-based tests. The findings are important because many universities, institutes and international testing organizations have implemented CBT during last decade. The findings of the current study suggest some recommendations for test developers and curriculum designers to investigate the comparability of computerized and paper-based exams and to measure the validity of tests before converting the PBT version of the test and introducing the CBT version as their current assessment tool.

6 Acknowledgement

We are grateful to the members of Language Department of Chabahar Maritime University (CMU) for their time, expertise and guidance integral to the completion of this piece of research work. We would like to give a heartfelt thank you to **Ms. *Vahide Shahbazi*** for all her sincere help, support and knowledge.

7 References

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

[2] Chapelle, C. A. (2008). Utilizing technology in language assessment. In E. Shohamy & N. H. Hornberger (Eds.). Encyclopaedia of language and education, 2nd Edition, Volume 7: language testing and assessment. https://doi.org/10.1007/978-0-387-30424-3_172.

[3] Lynch, R. (2000). Computer-based testing: The test of English as a foreign language (TOEFL). The Source, Fall 2000. Retrieved January 6, 2004, from http://www.usc.edu/dept/education/. The Source/>Fall2000.

[4] 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).

[5] Khoshsima, H., Hosseini, M. & Hashemi Toroujeni, S.M. (2017). Cross-Mode Comparability of Computer-Based Testing (CBT) versus Paper and Pencil-Based Testing (PBT): 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.

[6] Bugbee, A.C., & Bernt, F.M. (1990). Testing by computer: Findings in six years of use 1982-1988. Journal of Research on Computing in Education, 23(1), 87-100. https://doi.org/10.1080/08886504.1990.10781945.

[7] Olsen, J. B., Maynes, D. D., Slawson, D., & Ho, K. (1989). Comparison of paper-administered, computer-administered and computerized adaptive achievement tests. Journal of Educational Computing Research, 5, 311-326. https://doi.org/10.2190/86RK-76WN-VAJ0-PFA3.

[8] Jamieson, J. M. (2005). Trends in computer-based second language assessment. Annual Review of Applied Linguistics. 25, 228–242. https://doi.org/10.1017/s0267190505000127.

[9] Bennett, R. E. (2001). How the internet will help large-scale assessment reinvents itself. Education Policy Analysis Archives, 9. Retrieved June 18, 2005, from http://epaa.asu.edu/epaa/v9n5.html.

[10] Bennett, R. E. (2003). Online assessment and the comparability of score meaning. Princeton, NJ: Educational Testing Service. https://doi.org/10.14507/epaa.v9n5.2001.

[11] Khoshsima, H. & Hashemi Toroujeni, S.M. (2017d). A Comparative Study of the Government and Private Sectors' Effectiveness in ELT Program: A Case of Iranian Intermediate EFL Learners' Oral Proficiency Examination. Studies in English Language Teaching, Vol. 5, N. 1, February 2017; 86-108. ISSN 2372-9740 (Print), ISSN 2329-311X (Online). URL:http://dx.doi.org/10.22158/selt.v5nlp86.

[12] Khoshsima, H. & Hashemi Toroujeni, S.M. (2017a). Transitioning to an Alternative Assessment: Computer-Based Testing and Key Factors related to Testing Mode. European Journal of English Language Teaching, Vol.2, Issue.1, pp. 54-74, February (2017). ISSN 2501-7136. http://dx.doi.org/10.5281/zenodo.268576.

[13] Genc, H. (2012). An evaluation study of a call application: with belt or without belt. TOJET: The Turkish Online Journal of Educational Technology, 11(2). Retrieved July 2, 2011 from http://www.tojet.net/articles/v11i2/1125.pdf.

[14] OECD. (2010). PISA Computer-based assessment of student skills in science. http://www.oecd.org/publishing/corrigenda (accessed September 21, 2014). https://doi.org/10.1787/9789264082038-en.

[15] Chapelle, C. A. & Douglas, D. (2006). Assessing language to computer technology. Cambridge: Cambridge University press. https://doi.org/10.1017/CBO9780511733116.

[16] Lottridge, S., Nicewander, A., Schulz, M. & Mitzel, H. (2008). Comparability of Paper-based and Computer-based Tests: A Review of the Methodology. Pacific Metrics Corporation 585 Cannery Row, Suite 201 Monterey, California 93940.

[17] Noyes, J.M., Garland, K.J., and Robbins, E.L., (2004). Paper-based versus computer-based assessment: Is workload another test mode effect? British Journal of Educational Technology, 35, 111–113. https://doi.org/10.1111/j.1467-8535.2004.00373.x.

[18] Parshall, C.G., Spray, J.A., Kalohn, J.C., & Davey, T. (2002). Practical considerations in computer-based testing. New York, NY: Springer. https://doi.org/10.1007/978-1-4613-0083-0.

[19] Poggio, J., Glasnapp, D., Yang, X. & Poggio, A. (2005). A Comparative Evaluation of Score Results from Computerized and Paper & Pencil Mathematics Testing in a Large Scale State Assessment Program. The Journal of Technology, Learning and Assessment, 3(6), 5-30.

[20] 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.

[21] Association of Test Publishers. (2000). Computer-based testing guidelines. Washington, DC: Association of Test Publishers.

[22] International Test Commission. (2004). International Guidelines on Computer-Based and Internet- Delivered Testing. Retrieved January 21, 2011 from http://www.intestcom.org/itc_projects.htm. .

[23] American Council on Education. (1995). Guidelines for computerized-adaptive test development and use in education. Washington, DC: Author.

[24] International Test Commission. (2006). International guidelines on computer-based and Internet-delivered testing. International Journal of Testing, 6, 143–171. https://doi.org/10.1207/s15327574ijt0602_4.

[25] Pomplun, M., Frey, S., & Becker, D. F. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. Educational and Psychological Measurement, 62(2), 337-354. https://doi.org/10.1177/0013164402062002009.

[26] Bugbee Jr., A. C. (1996). The equivalence of paper-and-pencil and computer-based testing. Journal of Research on Computing in Education, 28, 282–299. https://doi.org/10.1080/08886504.1996.10782166.

[27] Khoshsima, H. & Hashemi Toroujeni, S.M. (2017b). Comparability of Computer-Based Testing and Paper-Based Testing: Testing Mode Effect, Testing Mode Order, Computer Attitudes and Testing Mode Preference. International Journal of Computer (IJC), (2017)

Volume 24, No 1, pp 80-99. ISSN 2307-4523 (Print & Online), http://ijcjournal.org/index.php/InternationalJournalOfComputer/article/view/825/4188.

[28] 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. .

[29] Flowers, C., Do-Hong, K., Lewis, P., & Davis, V. C. (2011). A comparison of omputer-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.

[30] 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.

[31] 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).

[32] Escudier, M., Newton, T., Cox, M., Reynolds, P., & Odell, E. (2011). University students' attainment and perceptions of computer delivered assessment: A comparison between computer-based and traditional tests in a 'high-stakes' examination. Journal of Computer Assisted Learning, 27(5), 440-447. https://doi.org/10.1111/j.1365-2729.2011.00409.x.

[33] Harde, P. L., Crowson, H. M., Xie, K., & Ly, C. (2007). Testing differential effects of computer-based, web-based, and paper-based administration of questionnaire research instruments. British Journal of Educational Technology, 30(1), 5-22. https://doi.org/10.1111/j.1467-8535.2006.00591.

[34] Sim, G., & Horton, M. (2005). Performance and attitude of children in computer based versus paper based testing. Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications (EDMEDIA) (pp. 3610-3614). Chesapeake, VA: AACE.

[35] 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, ISSN: 2289-3245. http://ijllalw.org/Current-Issue.html.

[36] 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.

[37] 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).

[38] Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. British Journal of Educational Technology, 33, 593-602. https://doi.org/10.1111/1467-8535.00294.

[39] Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The Effect of Computer- Based Tests on Racial-Ethnic and Gender Groups. Journal of Educational Measurement, 39(2), pp. 133-147. https://doi.org/10.1111/j.1745-3984.2002.tb01139.x.

[40] 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.

[41] Hosseini, M., Zainol Abidin, M. J., Baghdarnia, M., (2014). Comparability of Test Results of Computer-Based Tests (CBT) and Paper and Pencil Tests (PBT) among English Language Learners in Iran. International Conference on Current Trends in ELT, 659-667. https://doi.org/10.1016/j.sbspro.2014.03.465.
[42] Kauffman, A. S. (2003). Practice effects. Speech and Language Forum. http://www.speechandlanguage.com/cafe/13.asp.
[43] Bachman, L. F., & Palmer, A. S. (1996). Language testing in practice. Oxford, England: Oxford University Press.
[44] Warner, R. M. (2013). Applied Statistics: From Bivariate through Multivariate Techniques. (2th Ed.). SUA: SAGE Publication Inc.
[45] Creswell, J. w. & Clark, V. (2007). Designing and Conducting Mixed Methods Research. London: Sage Publications.
[46] Larson-Hall, Jenifer. (2010). A guide to doing statistics in second language research using SPSS. New York: Routledge.
[47] 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.
[48] Al-Amri, S. (2007). Computer-based vs. Paper-based Testing: Does the test administration mode matter. Proceedings of the BAAL Conference 2007.
[49] Seidman, I. (1998). Interviewing as qualitative research: A guide for researchers in education and the social sciences (2nd Ed.). New York: Teachers College Press.
[50] Al-Amri, S. (2008). Computer-based testing vs. paper-based testing: A comprehensive approach to examining the comparability of testing modes. Essex Graduate Student Papers in Language and Linguistics, 10, 22–44. Retrieved January 28, 2012 from http://www.essex.ac.uk/linguistics/publications/egspll/volume_10/pdf/EGSPLL10_2244SAA_web.pdf.

8 Authors

Hooshang Khoshsima is an associate professor at Language Department, Chabahar Maritime University. He has published a number of textbooks and many research papers in several journals and conferences. His areas of interests are research in translation issues, applied linguistics, teaching methodologies, assessment, testing, and ESP.

Seyyed Morteza Hashemi Toroujeni was born in Mazandaran, Iran, in 1983. He received his B.A. in English Language Translation from the PNU in 2007, and his M.A. in TEFL from the CMU in 2016. He is passionate about alternative assessment, educational technology, and psychometrics and how they can be used to improve different aspects of our education. He is the editorial board and reviewer of some international journals such as iJET, AJHC, IJLA & IJEAP. He also teaches university courses in TEFL.

Article submitted 11 March 2017. Published as resubmitted by the authors 15 June 2017.