COMPUTER-BASED (CBT) VS. PAPER-BASED (PBT) TESTING: MODE EFFECT, RELATIONSHIP BETWEEN COMPUTER FAMILIARITY, ATTITUDES, AVERSION AND MODE PREFERENCE WITH CBT TEST SCORES IN AN ASIAN PRIVATE EFL CONTEXT

by **Hooshang Khoshsima,** Chabahar Maritime University, Chabahar, Iran,
khoshsima2002@yahoo.com
Seyyed Morteza Hashemi Toroujeni, Chabahar Maritime University, Chabahar, Iran,
hashemi.seyyedmorteza @ gmail.com
Nathan Thompson, Assessment Systems Corporation, USA, nthompson @ assess.com
and **Mohammad Reza Ebrahimi,** Gonabad University, Gonabad, Iran,
rezaebrahimi2@yahoo.com

Abstract

The current study was conducted to investigate whether test scores of Iranian English as Foreign Language (EFL) learners were equivalent across CBT and PBT modes, with 58 intermediate learners studying at a private language academy located in Behshahr city in northern Iran. Moreover, test takers' computer familiarity, attitudes, aversion, and testing mode preference were regarded as the potential issues to influence CBT test scores. Data were collected using CBT and PBT versions of Nelson Proficiency Multiple-Choice Tests and Computer Aversion, Attitudes, and Familiarity Index (CAAFI) questionnaire as well as a simple testing administration mode preference question. The participants produced similar scores across modes, although they insignificantly outperformed on the CBT version. Additionally, analysis of the overall scores on the CAAFI and mode preference question obtained from CBT testing session indicated no statistically significant correlation between computer familiarity, attitude, aversion, and mode preference variables and test takers' CBT scores. The qualitative findings of this study obtained by semi-structured interview revealed that most of the participants showed high preference and more advantages for CBT over PBT to rationalize why they preferred this mode of testing.

Keywords: Computer-Based Testing; testing administration mode; computer familiarity, attitudes and aversion; testing administration mode preference

1. Introduction

In the last decades, computer technology and related technological tools have been extensively utilized in language testing to analyze scores and results quickly (Boeve et al., 2015; Laborda & Penalver, 2018). According to Daniels and Gierl (2017), computer-based testing (CBT) brings many benefits in educational contexts. Students are provided with positive interactions or communication opportunities and can receive immediate feedback (Daniels and Gierl, 2017). Moreover, it is cost-effective, and the availability of powerful computers in educational contexts make computer-based test delivery both feasible and attractive (Boeve et al., 2015). CBT also provides test takers with the opportunity of taking their tests at any time and place. The introduction of IBM model 805 scoring machine in Western countries was recorded as the first use of the computer in language testing in 1935; yet, its prevalence in educational assessment domain is rather slow (Boeve et al., 2015) especially in Asian developing countries. The causes shall be attributed to some barriers such as limited access to computers and concern of the effects of the transition from paper to computer on test takers' scores that is commonly defined as "testing administration mode" effect.

Testing administration mode effect is the main concern of Asian researchers from the countries such as Iran, Turkey, China, Malaysia, Saudi Arabia, and Jordan when they begin to implement CBT along with PBT in their educational system or consider CBT to replace PBT. Then, they investigate whether test takers' scores are equivalent across two modes (e.g., Chen et al., 2014; Khoshsima & Hashemi Toroujeni, 2017a; Alakyleh, 2018; Yurdabakan and Uzunkavak, 2012). Equivalency or interchangeability of scores from CBT and PBT has been a controversial issue during the last decade (Sangmeister, 2017). How changing the administration mode can affect students' test performance is a crucial question when considering changing from PBT to CBT. Furthermore, the interaction between individual differences (e.g., prior computer familiarity, attitudes, and aversion) and CBT performance should be investigated in equivalence studies in which the score equivalency and reliability are examined to replace CBT with PBT.

Since growing concerns over the impact of computer familiarity, attitudes, aversion and testing mode preference on EFL attainments of the private sector from CBT exist, the current research aimed to investigate the equivalency of CBT and PBT and address testing administration mode effect on test takers' scores by discovering similarities or differences between the mean scores of CBT and PBT versions of a test. It was conducted to help to accelerate the move to CBT due to all its benefits mentioned.

2. Literature review

As the relevant literature is reviewed, the empirical evidence shows that two identical CBT and PBT (Paper-Based Testing) do not always result in the same scores. Hence, these conclusions are referred to as “testing mode effect”: the effects of the transition from paper to computer on performance in two similar or equivalent tests. International Guidelines on CBT state that when a test is implemented in two modes and two sets of similar scores are obtained, the scores are considered equivalent and reliable (ITC, 2016). The equivalent test scores established for two CBT and PBT modes (AERA, 2014) demonstrate that computer-based testing is valid and reliable. Based on the classical True-Score Theory, the same test implemented in two modes, i.e., CBT and PBT, should result in equivalent or identical test scores. The transition from paper to computer took place long ago in Westernized or heavily Westernized countries, but in many countries such as Asian developing countries, it has not happened yet because computer and internet access is limited. Then, developing CBTs must be done with utmost care, due to limited access to the internet in Asian developing countries. Mangen et al. (2013) investigated the impact of test version (CBT and PBT) on test achievements of 72 students. Their findings showed a great difference between CBT and PBT performances. The students gained significantly higher scores in CBT format of the test (Mangen et al., 2013). In one of the recent equivalence studies done by Washburn et al. (2017), the performance and perception of CBT vs. PBT were evaluated concerning the transitioning from traditional paper-based to CBT. The findings of the study showed that the students’ scores for the CBT version of the test were higher than those obtained from the PBT version (Washburn et al., 2017). Moreover, it is recommended to eliminate the possible effects of moderator variables such as computer familiarity (Jeong, 2014), attitudes toward the use of computer (Dammas, 2016), computer aversion (Balogun & Olanrewaju, 2016) and mode preference (Boeve et al, 2015; Mizrachi, 2015) on test scores.

2.1. Computer familiarity and attitude

There is a difference in the test takers’ familiarity with the computer. It seems that EFL learners who are frequent users of computers and the internet and are more familiar with computers attain dramatic educational gains on CBT. Kirsch, Jamieson, Taylor, and Eignor’s (1998) research findings on computer experience and CBT performance on a TOEFL test (after implementation of online familiarization training) showed no significant relationship between prior computer use of test takers and their performance on the computerized test.

Computer attitudes or prior attitudes toward the use of computer play a crucial role in implementing CBT successfully. Some studies indicate that test takers have positive attitudes toward CBT (Al-Amri, 2009). In another study by Al-Amri (2009) using some sections of the CAS questionnaire to study learners' attitudes toward computer use, he reported that students showed a high preference for CBT, although no relationship between learners' attitudes and their performance on CBT was detected. Youdbakan and Uzunkavak (2012) reported a study investigating learners' attitudes toward computer and CBT among 784 Turkish primary school learners in private and state schools using a researcher-constructed attitude scale.

However, even though, based on conclusive evidence of a higher education context, Khoshsima & Hashemi Toroujeni (2017b) claimed that moderator variables such as computer attitudes and mode preference are not considered factors that might affect students' performance on CBT, many Asian test users and test developers are not optimistic about the generalizability of the findings to the private EFL sector.

2.2. Computer aversion and testing mode preference

McDonald (2002) reported that computer aversion is an unpleasant feeling of fear and uneasiness experienced by a student when s/he is working with a computer. According to McDonald (2002), the actual effects of computer aversion (sometimes called computer anxiety) on test takers' performance on CBT is not clear and conclusively definite. However, test takers who have a strong aversion toward the use of computer experience achieve low performance in CBT (Balogun & Olanrewaju, 2016).

To examine the relationship between test takers' preference and their test scores, the researchers use either preference scale questionnaire or interviews to ask which testing mode of administration they prefer (e.g., Al-Amri, 2009; Corlett-Rivera & Hackman, 2014; Mizrachi, 2015). In a study done by Al-Amri (2009), although test takers preferred to take CBT, their test performance was better on PBT.

In the current study, individual differences or characteristics are considered of great importance and it was hypothesized that there was no statistically significant difference between the mean of two sets of scores obtained from CBT and PBT. Also, the correlations between computer familiarity, computer attitudes, computer aversion and testing mode preference with test performance were also investigated based on the hypotheses that there was no statistically significant impact of the participants' level of computer familiarity, attitudes, aversion and preference toward computers on their test performance using CBT. Then, considering the above discussion, it is necessary to investigate testing administration mode effect, the relationship of computer familiarity, attitudes, and aversion toward computers with the performance of test takers on their CBT test scores. The results of the study could inform testing practitioners when designing testing in private EFL contexts.

3. The current study

3.1. Objectives of the study

Since evaluating the equivalency or comparability of PBT and CBT tests is crucial before introducing CBT into any context, the following research questions were investigated:

RQ1. Is there a significant difference in test scores for CBT and PBT testing modes?

RQ2**.** Do participants' computer familiarity, computer attitude, computer aversion, and testing mode preference affect test scores using CBT?

Then, to investigate the problems raised by the study the following null hypotheses will be addressed.

H0 1: There is no statistically significant difference in CBT and PPT test scores among Adrina Language Academy (ALA) EFL Learners.

H02: Participants' computer familiarity, computer attitude, computer aversion, and testing mode preference do not affect test scores using CBT.

3.2. Participants

This study was carried out in autumn 2017 at the Adrina Language Academy (ALA) located in Behshahr city, in northern Iran, Mazandaran province. 108 English as Foreign Language (EFL) adult learners who were taking the General English Courses of different levels at ALA took the TOEFL general proficiency test (Phillips, 2001) (PBT Complete Test/p.515-538) as a reliable and valid index of general English proficiency for organizing a homogenous testing group in Summer 2017. Based on the general English language proficiency conversion table, 58 intermediate EFL learners (the overall TOEFL score ranged from 477 to 510) were selected as homogenous ones to participate in the main investigation. The 58 participants consisted of 30 males (51.72%) and 28 females (48.28%). The age range of the 58 students was between 18 to 34 years with a mean of 23.9 years.

Students who were participating in the study were given a consent form to sign. The subjects were told that their responses to tests and questionnaire would be anonymous and that the results would be used for research purposes only.

3.3. Design, instrumentation and procedure

The present study consisted of three sub-studies. The first study used two CBT, and PBT versions of two equivalent tests was to examine the effect of testing administration mode on test scores to answer research question one. The second study used a questionnaire and preference question was designed to investigate the relationship of computer familiarity, attitude, aversion and mode preference with CBT test scores to answer Research Question 2. The third study consisted of the interview as a qualitative instrument to inquire about participants' testing administration mode preference, attitudes toward PPT and CBT, development of positive or negative attitudes and their opinions about two test versions. The learners were assigned to one testing group based on *common person design* (repeated measures or pre and post-test design).

The quantitative data collected from the questionnaire could not access the unexpected reasons why test takers had particular perceptions of various aspects of the tests (CBT and PBT) they took. Hence, subsequent interview (qualitative data) was used to allow test takers to explain their reasons in their voice (Research Question 2 related to attitudes towards the use of the computer in CBT test condition and testing mode preference).

The multiple choice achievement tests used in the PBT and CBT versions were from the Nelson Proficiency Tests (Test 200A and Test 200B for intermediate level students) selected from Nelson English Language Tests by Fowler and Coe (1976). The battery consists of 40 separate tests, 4 tests of which are equivalent in difficulty at each of 10 levels from beginners to advanced. Financial considerations and practical ones discouraged us from adopting a newer version which may not be necessarily different (as there is a need for doing a pilot study in advance). These two equivalent tests were used to mitigate possible testing effects caused by using the same test on two occasions. Test 200A was used as the PBT version of the test and Test 200B as the equivalent test was converted into the CBT version. These standard tests included fifty multiple-choice items to assess the grammatical knowledge and structural progression of the participants. The 40 Nelson English Language Tests were designed independently and are appropriate for ten different levels of language proficiency. The tests were designed for a passing mark of 30 (60%).

To convert the PBT version of the test (Test 200B) into its CBT counterpart, a professional web-based testing service provided by Classmarker.com was used. The identical tests were used in both PBT and CBT for pairwise comparison because this design needs a smaller research sample (Sangmeister, 2017). In CBT session, each test taker was given a registration code to activate his/her testing account and to enter the testing environment. Each test taker was given a computer and s/he should answer the questions appearing on the screen one by one. The clear and straightforward question: *Which one would you prefer? Taking the test on paper – no difference – on computer screen* appeared on the screen at the end of the CBT test to get correct feedback on the correlation of preference towards administration mode with test takers scores.

Another research instrument was used to measure computer aversion, computer attitudes, and computer familiarity. The questionnaire was based on the revised version of Computer Aversion, Attitudes, and Familiarity Index (CAAFI) by Schulenberg and Melton (2008). According to Hashemi (2016), the CAAFI is a powerful instrument to gain a good understanding of these constructs. This 30-item questionnaire was composed of three factors: factor 1 was related to the computer familiarity construct with items 3, 13, 14, 16, 20-23, 27, and 30, factor 2 was related to the computer attitudes construct with items 1, 2, 4, 5, 8, 11, 18, 19, 28, and 29, and factor 3 was related to the computer aversion construct with items 6, 7, 9, 10, 12, 15, 17, and 24-26. The factor structure of CAAFI had been confirmed using confirmatory factor analysis procedure and analysis of internal consistency reliability coefficients (Schulenberg & Melton, 2008). In addition to the exploratory and confirmatory factor analysis, enough details on the primary development of this questionnaire were provided by Schulenberg (2002), Schulenberg, Yutrzenka and Gkhm (2006) and Schulenberg and Melton (2008). The items had a seven-point scale from -3 (absolutely false) to 3 (absolutely true) to increase the response rate. Zero, in this range, shows a neutral response toward an individual statement. In this questionnaire, some of the statements are negatively worded that necessitate reverse scoring. The negatively worded items 6, 8, 9, 15, 17, 24, 25, and 26 should be reverse scored. For each factor, the items were summed, and higher positive scores suggested less computer anxiety, more positive or favorable attitudes toward the computer and more experience and familiarity with computers.

Based on the descriptive data, computer familiarity factor had a mean of 14.39 (SD=8.54), and α of .846. Computer attitudes had a mean of 6.50 (SD=8.13) and α of .664, and Computer aversion had a mean of 9.05 (SD=10.55) and α of .855. The CAAFI had an overall α of .906 and a mean of 29.94 (SD=24.44). Therefore, the internal consistency reliability and descriptive results obtained in the present study were comparable with the findings provided by Schulenberg, Yutrzenka and Gohm (2006), and Schulenberg and Melton (2008). The means for the three factors in the CAAFI questionnaire were obtained by summing the responses of respondents on ten items (Likert-rating scale of 7) measuring each factor. The questionnaire also collected data on the participants' demographic information such as name, age, and level of education. Cronbach's α reliability analysis was performed as a measure of internal consistency for the CAAFI questionnaire in this study, and a high-reliability coefficient of α=.906 was achieved for the 30 items CAAFI index.

A set of predetermined open-ended questions (Appendix A) were asked to 26 randomly selected participants as a semi-structured interview to inquire about their testing administration mode preference, attitudes toward PPT and CBT, development of positive or negative attitudes and their opinions about the features of two test versions. The researchers were interested in using a semi-structured interview because questions could be prepared in advance and the interviewees could express themselves easily in the ways they preferred. The questions of the interview were developed by the researchers and then content was analyzed by two experts of TEFL. This qualitative method was used to support the quantitative research data.

Both quantitative and qualitative methods were used to collect data to answer the research questions of the study and confirm or reject the research null hypotheses. After the TOEFL placement test, 58 students at the intermediate level were chosen as the sample. The participants took the Nelson Test 200A as the PBT version of the test on the first testing occasion (50 questions in 50 minutes). To eliminate testing effects, after a three-day interval, the same participants took the equivalent Nelson Test 200B in CBT version (50 questions in 50 minutes). After completing the CBT, the testing mode preference question appeared on the screen. Then, the CAAFI questionnaire was distributed to the participants. Also, 26 randomly selected participants of the study were interviewed for 7-10 minutes after the CBT session.

4. Results

Kolmogorov-Smirnov test indicated that PBT scores and CBT scores significantly deviated from normality (Table 1), then, the nonparametric Wilcoxon signed-rank test equivalent of the paired samples t-test was chosen to compare the test scores on the PBT and CBT versions.

Table 1. Results of Normality tests for PBT and CBT versions

One-Sample Kolmogorov-Smirnov Test

		PBT	CBT
N		58	58
	Mean	43.72	45.46
	Std. Deviation	7.78	4.38
Most Extreme Differences	Absolute	.186	.184
	Positive	.141	.151
	Negative	-.186	-.184
Kolmogorov-Smirnov Z		1.41	1.40
Asymp. Sig. (2-tailed)		.036	.040

Wilcoxon signed-rank test was used to measure changes in the ranked positions of PBT and CBT scores for the 58 participants and provide the differences in ranked data between the CBT and PBT test scores including the mean rank and sum of ranks. As evidenced in Table 2, 25 participants received higher scores in PBT session than in CBT (negative ranks showed the ranks for which the PBT scores were higher than the CBT scores), and 29 participants received higher scores in CBT session than in PBT. Another four participants experienced no difference in their scores in the two test conditions.

Table 2. Rank-based descriptive statistics of testing sessions

Ranks

		N	Mean Rank	Sum of Ranks
CBT - PBT	Negative Ranks	25[a]	24.42	610.50
	Positive Ranks	29[b]	30.16	874.50
	Ties	4[c]		
	Total	58		

a. CBT < PBT / b. CBT > PBT / c. CBT = PBT

The results of Wilcoxon signed-rank test indicated that test scores were not significantly different for the two test modes (CBT vs. PBT) (Z = -1.137, p = 0.255). Since the PBT and CBT test scores (Table 1) and the scores for computer attitudes, computer aversion and computer familiarity (Table 3) were not normally distributed, Spearman’s rank-order correlation analyses were used to investigate the relationships between computer familiarity, attitudes, aversion, and CBT test scores.

Table 3. Results of normality tests for each factor of CAAFI

One-Sample Kolmogorov-Smirnov Test

		Computer familiarity	Computer attitudes	Computer aversion
n		58	58	58
	Mean	14.39	6.5	9.05
	Std. Deviation	8.54	8.13	10.55
Most Extreme Differences	Absolute	.28	.25	.24
	Positive	.28	.25	.24
	Negative	-.22	-.17	-.14
Kolmogorov-Smirnov Z		2.15	1.93	1.85
Asymp. Sig. (2-tailed)		.00	.00	.00

Spearman’s rank-order correlation results showed that the null hypothesis was not rejected and there was no statistically significant correlation between CBT test scores and computer familiarity (r (56) =.182, p=.172). The results of Spearman’s rank-order correlation test also showed that there was no statistically significant relationship between computer attitudes and CBT test scores (r (56) =.094, p=.483) and the null hypothesis was not rejected. However, there was a statistically significant relationship between computer aversion and CBT test scores (r (56) =.287, p=.029). The null hypothesis was rejected. As can be concluded from the results, there was no significant correlation between computer familiarity and attitudes toward computer and CBT test scores. Findings of the current study on the relationship between computer familiarity and CBT test scores were in line with the findings of studies such as Jeong (2014), who found no relationship between the two variables.

Spearman's rank-order correlation analysis for 58 test takers' testing mode preference and their CBT performance showed no statistically significant correlation (r (56) =.203, p=.127). Then, the null hypothesis for testing mode preference was confirmed based on the evidence that this variable was not a statistically significant predictor of CBT scores. Additionally, there was no statistically significant correlation between testing mode preference and PBT test scores (r (56) =-.069, p=.607).

Since the data normality assumption of dependent variable was violated, and the scores came from the same test takers, Wilcoxon signed-rank test was used to compare both PBT and CBT mean rank of three mode preference groups (coded as 1=PBT, 2=No-Difference, 3=CBT based on the testing mode preference question). The comparison was made to examine the effect of testing mode preference on their performance and whether test takers outperformed in their preferred testing mode session. Out of 58 test takers who answered the preference question, 32 preferred taking CBT (55%), 18 preferred taking PBT (31%). 8 (14%) didn't mind taking the test on either mode.

Wilcoxon signed-rank test demonstrated that the median CBT ranks for PBT mode preference group, Mdn=47, were not statistically significantly higher than the median PBT ranks, Mdn=48, Z=-.491, P=.624. It meant that although those test takers who preferred to take the test in the PBT version performed slightly better in their PBT session, there was no statistically significant difference between their PBT and CBT test scores. The same results were attained for the other two No-Difference, and CBT mode preference groups and the median CBT test ranks of two preference groups were not statistically significantly higher than the median PBT test ranks; PBT Mdn=47 vs CBT Mdn=50, Z=-1.633, p=102 and PBT Mdn=45 vs CBT Mdn=45, Z=-.405, p=.686 for No-Difference and CBT mode preference groups, respectively. The results show that 55 % of the test takers who preferred taking the test on CBT (CBT mode preference group) did the same on two PBT and CBT versions of the tests. It was concluded that although the test takers preferred to take the CBT version of the test, they did not outperform in their preferred mode and there was no statistically significant difference in their test scores received from two PBT and CBT versions.

Subsequently, a semi-structured interview was conducted and responses from the open-ended questions were transcribed. Content analysis was conducted on the transcribed data by identifying the main concepts using thematic analysis. Based on the results and findings from the interview data, of the 26 participants interviewed, 18 (69%) favored CBT and 8 (31%) preferred PBT. They were then asked about the features of two test versions they preferred and didn't prefer, about their testing administration mode preference before implementing PBT and after administering CBT as well as their reasons behind their preferences and mode preference change (in the case of changing mode preference).

Those who advocated CBT mentioned fifteen positive features. All the 18 interviewees who favored CBT stated that they could easily read the test items on a computer screen, choose and change answers, and obtain immediate feedback or test scoring reports. Eleven (61%) of the 18 interviewees stated that they liked the CBT testing environment because they could read one question on each page, they should click to highlight the correct answer, and they were able to see the time on the corner of the screen. Eight (45%] of CBT advocators found the CBT version to be a less fatiguing and more enjoyable test environment due to certain elements of the screen such as colors, graphics, and text together. Furthermore, nine (50%), sixteen (90%], and twelve (66%] of these 18 interviewees were of the opinions that the CBT was a more comfortable and faster-testing mode, with fewer response recognition errors. They believed that they could recognize the correct answer among the options easily. Out of these 18 interviewees who favored CBT, four (22%] of them stated that the CBT needed less time to review the question items and modify answers, and it took less time to respond to the questions. Fourteen (78%], eleven (61%], and ten (55%] of these interviewees also commented on enhanced security, faster decision making as a result of immediate scoring and score reporting, and causing less stress and anxiety of CBT, respectively. Furthermore, five (30%] of them commented on the accuracy in CBT while sixteen (90%] felt that CBT eliminated the human error in scoring and improved the quality and reliability of the test. CBT advocators stated that they didn't prefer PBT because it was boring but taking the test on the computer was like a game.

Out of the 26 interviewees, twenty-three (88%) asserted that they did not have to use their hands to write answers or check the correct answer on the paper. They stated that this feature makes taking CBT easier. Although four (15%] of the interviewees reported that they had a problem with the mouse when it stopped working for some seconds, they still liked the

CBT version. However, eight interviewees did not prefer CBT over PBT. All the eight respondents (100%) who preferred to take PBT stated that they could write down or underline some key-words or phrases for future returning. In PBT, they could put a bullet next to the questions they did not know their answers for future review. Five (62%) claimed that CBT required more technical knowledge. Six (75%] also expressed their concern of system breaking down and crash. They were afraid of computers not working as they expect during the test. Seven of these eight interviewees felt that reviewing the answers in CBT was time-consuming (87%). Three (37%] of PBT advocators commented on the challenges caused by scrolling horizontally or vertically on some long pages such as score reporting page. Concerning the testing mode preference change, nine of the interviewees (35%] stated that they changed their preference in favor of CBT after taking this version. They declared that they had never taken CBT test and they did not mind taking the test in either mode, but after benefiting from CBT in the second testing session, they had positive attitudes toward it and preferred taking this version in the future.

Then, the number of test takers who opted for CBT increased by 27% after taking the test. According to the results, it was concluded that the number of participants who preferred PBT or did not mind taking the test in either mode before taking CBT changed in favor of the test takers who chose CBT as their preferred testing mode preference after taking CBT. Surprisingly, all of them stated that they became positive toward CBT due to receiving immediate feedback and test results and allowing them to see if they passed the exam.

5. Discussion

The fact that no statistically significant difference in test scores for the participants of this study who took the PBT and CBT equivalent tests existed suggests that the two modes can represent grammatical competence validly and reliably, and CBT does not have a significant effect on test takers' scores.

Based on the findings, the concern of the differential CBT test scores due to prior familiarity with a computer is eliminated. It may be claimed that as the learners of the current decade are fully familiar with a computer through playing games or using the internet and communicating via different kinds of messengers, computer familiarity is losing its importance and relationship with CBT performance. The lack of variance in PBT and CBT scores in the present study and some other studies may be the effect of generational difference; the present generation is more familiar with technology and has more exposure to it. No correlation between attitudes toward the use of computer and CBT scores suggests that this variable may not be considered as a source of variance in PBT and CBT performances. Findings of the current study were consistent with the results reported in Al-Amri (2009), who found no statistically significant correlation between computer attitudes and CBT performance and concluded that test takers' attitudes (either positive or negative) did not affect their CBT performance.

According to the observational results of the study done by Labora and Penalver, test aversion still seems to be a critical issue, in spite of the new generation's familiarity with new technologies like a computer (Laborda & Penalver, 2018). Mastuti and Handoyo (2017) stated that aversion towards the implementation of CBT is still well worthy of investigation. The current study showed a weak positive relationship between computer aversion and CBT test scores. As higher scores on computer aversion items indicated less computer aversion, the positive correlation between computer aversion and CBT test scores indicated that less anxiety toward the use of the computer would lead to higher scores on CBT or vice versa. Also, Spearman's rank-order correlation test was run to look at the relationship of testing mode preference and CBT scores. The results indicated no association between mode preference and CBT score. The comparison of PBT and CBT scores of mode preference groups (those who preferred the PBT version and those who preferred the CBT version) revealed that in spite of the preference for PBT and CBT versions, there was no significant difference between the scores obtained from each test version and test takers did not perform better in their preferred mode. Those participants who preferred taking PBT did the same in their CBT exam.

Additionally, those who preferred taking the CBT test did not outperform the PBT ones in their exam. Accordingly, based on the Wilcoxon signed-rank test, no statistically significant difference was found between the PBT and CBT performance of preference groups and their preferred test mode performances. Furthermore, those who did not mind taking the test in either mode did better in CBT, but the difference was not statistically significant. The results suggest that the mode preference and eagerness of test takers do not validate a CBT test, and the standard guidelines for establishing equivalence between PBT and CBT should be followed.

As evidenced by the quantitative part of the study, most test takers preferred to take the CBT version of the test. Among the interviewees, 69% of them declared that they preferred to take the test in the CBT version. The qualitative findings supported the quantitative results.

6. Conclusions, recommendations and limitations

Based on the findings, it is argued that teachers and test developers may invest in spreading CBT through private EFL contexts and motivate learners to take it. Language teachers should give their learners more opportunities to begin working with computer and CBT version in classes and keep in mind that CBT may be especially appealing to the present generation of learners who are growing up with technology and computers.

Since the research indicated that students feel quite comfortable with taking the CBT version of the test and prefer this kind of testing (Khoshsima & Hashemi Toroujeni, 2017h), it can be used as an alternative assessment instrument in private EFL contexts. However, the findings of the current study cannot be generalized to all contexts and participants with different background of knowledge or field of study. Since only intermediate Persian English as Foreign Language Learners of a private institution participated in this research, further studies with more heterogeneous participants (with different educational background, level of English proficiency, nationality and ethnicity) are needed to increase generalizability over time with different tasks or tests.

References

Alakyleh, A. S. (2018). Evaluating the comparability of (PPT) (CBT) by implementing the compulsory Islamic Culture Course Test in the University of Jordan. *International Journal of Assessment Tools in Education, 5*(1), 176-186. DOI: 10.21449/ijate.370494.

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.* Unpublished Doctor of Philosophy in Linguistics thesis. Colchester: University of Essex.

American Educational Research Association, (2014). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Balogun, A. G., & Olanrewaju, A. S. (2016). Role of computer self-efficacy and gender in computer-based test anxiety among undergraduates in Nigeria. *Psychological Thought*, *9*(1), 58-66. https://doi.org/10.5964/psyct.v9i1.160

Boevé, A. J., Meijer, R. R, Albers, C. J, Beetsma, Y., & Bosker, R. J. (2015). Introducing computer-based testing in high-stakes exams in higher education: results of a field experiment. *PLOS ONE*, *10*(12) e0143616. DOI:10.1371/journal.pone.0143616.

Chen, G., Cheng, W., Chang, T.-W., Zheng, X., & Huang, R. (2014). A comparison of reading comprehension across the paper, computer screens, and tablets: Does tablet familiarity matter? *Journal of Computers in Education, 1*(2-3), 213-225.

Corlett-Rivera, K., & Hackman, T. (2014). E-book usage and attitudes in the humanities, social sciences, and education. *Portal: Libraries and the Academy*, *14*(2), 255-286.

Dammas, A. H. (2016). Investigate students' attitudes toward the Computer Based Test (CBT) at chemistry course. *Archives of Business Research, 4*(6), 58-71.

Daniels, L. M., & Gierl, M. J. (2017). The impact of immediate test score reporting on university students' achievement emotions in the context of computer-based multiple-choice exams. *Learning, and Instruction*, 52, 27-35. http://dx.doi.org/10.1016/j.learninstruc.2017.04.001

Fowler, W. A., & Coe, N. (1976). *Nelson English Language Tests*. Ontario: Thomas Nelson and Sons Ltd.

García Laborda, J., & Alcalde Penalver, E. (2018). Constraining issues in face-to-face and Internet-based language testing. *Journal for Educators, Teachers, and Trainers, 9*(2), 47-56.

The International Test Commission (2006) International Guidelines on Computer-Based and Internet-Delivered Testing, *International Journal of Testing*, *6*(2), 143-171, DOI: 10.1207/s15327574ijt0602_4 http://dx.doi.org/10.1207/s15327574ijt0602_4.

Jeong, H. (2014). A comparative study of scores on computer-based tests and paper-based tests. *Behavior & Information Technology*, *33*(4), 410-422.

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, 2*(1), 54-74. http://dx.doi.org/10.5281/zenodo.268576.

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)*, *24*(1), 80-99. http://ijcjournal.org/index.php/InternationalJournalOfComputer/article/view/825/4188.

Khoshsima, H., & Hashemi Toroujeni, 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)*, *12*(10). https://doi.org/10.3991/ijet.v12i10.6875.

Kirsch, I., Jamieson, J., Taylor, C., & Eignor, D. (1998). *Computer Familiarity among TOEFL Test Takers* (TOEFL Research Report 59). Princeton, NJ: Educational Testing Service.

Mangen, A., Bente, R. W., & Kolbjørn, B. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research*, *58*, 61-68.

Mastuti, E., & Handoyo, S. (2017). Effects of individual differences on the performance in computer-based test (CBT). *Advances in Social Science, Education, and Humanities Research: Proceedings of the 3rd ASEAN Conference on Psychology, Counselling, and Humanities (ACPCH, 2017)*. Malang: University of Muhammadiyah. http://dx.doi.org/10.2991/acpch-17.2018.44.

McDonald, A. S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers & Education*, *39*(3), 299-312. https://doi.org/10.1016/S0360-1315(02)00032-5.

Mizrachi, D. (2015). Undergraduates' academic reading format preferences and behaviors. *The Journal of Academic Librarianship*, *41*(3), 301-311. http://dx.doi.org/10.1016/j.acalib.2015.03.009.

Phillips, D. (2001). *Longman Complete Course for the TOEFL Test: Preparation for the Computer and Paper Tests*. White Plains, NY: Longman.

Sangmeister, J. (2017). Commercial competence: Comparing test results of paper-and-pencil versus computer-based assessments. *Empirical Research in Vocational Education and Training*, *9*(3). DOI: 10.1186/s40461-017-0047-2.

Schulenberg, S. E. (2002). The development of computer aversion, attitudes, and familiarity index (CAAFI). *Dissertation Abstracts International*, *62*(12), 5978B, (UMI No. 3037841).

Schulenberg, S. E., & Melton, A. M. A. (2008). The computer aversion, attitudes, and familiarity index (CAAFI): A validity study. *Computers in Human Behavior*, *24*(6), 2620-2638. https://doi.org/10.1016/j.chb.2008.03.002.

Schulenberg, S. E., Yutrzenka, B. A., & Gohm, C. L. (2006). The computer aversion, attitudes, and familiarity index (CAAFI): A measure for the study of computer-related constructs. *Journal of Educational Computing Research*, *34*(2), 129-146. https://doi.org/10.2190/45B4-GMH7-GEQB-T1H1.

Washburn, S., Herman, J., & Stewart, R. (2017). Evaluation of performance and perceptions of electronic vs. multiple-choice paper exams. *Advances in Physiology Education*, *41*(4), 548-555. DOI: 10.1152/advan.00138.2016.

Yurdabakan, I., & Uzunkavak, C. (2012). Primary school students' attitudes toward computer-based testing and assessment in turkey. *Turkish Online Journal of Distance Education*, *13*(3), 177-188.

Appendix 1. Semi-structured interview questions

1.	Which mode of testing administration did you prefer?
2.	Which features of the paper-based test did you prefer?
3.	Which features of paper-based test didn't you prefer?
4.	Which features of computer-based test did you prefer?
5.	Which features of computer-based test didn't you prefer?
6.	What was your testing administration mode preference choice before taking paper-based testing?
7.	What was your testing administration mode preference choice after taking computer-based testing?
8.	(In the case of changing mode preference) what was/were the reason(s) that you changed your mode preference choice?