English Language Teaching; Vol. 10, No. 2; 2017
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

Cross-Mode Comparability of Computer-Based Testing (CBT) Versus Paper-Pencil Based Testing (PPT): An Investigation of Testing Administration Mode among Iranian Intermediate EFL Learners

Hooshang Khoshsima[1], Monirosadat Hosseini[2] & Seyyed Morteza Hashemi Toroujeni[1]

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

[2] Department of English Language, Faculty of Humanities, Tehran Payame Noor University, Iran

Correspondence: Seyyed Morteza Hashemi Toroujeni, Foreign Languages Department, Faculty of Management and Humanities, Chabahar Maritime University, Iran. Tel: 98-911-789-7630. E-mail: Hashemi.seyyedmorteza@gmail.com & M.hashemi@cmu.ac.ir

Received: November 7, 2016 Accepted: January 3, 2017 Online Published: January 5, 2017

doi: 10.5539/elt.v10n2p23 URL: http://dx.doi.org/10.5539/elt.v10n2p23

Abstract

Advent of technology has caused growing interest in using computers to convert conventional paper and pencil-based testing (Henceforth PPT) into Computer-based testing (Henceforth CBT) in the field of education during last decades. This constant promulgation of computers to reshape the conventional tests into computerized format permeated the language assessment field in recent years. But, enjoying advantages of computers in language assessment raise the concerns of the effects that computerized mode of testing may have on CBT performance. Thus, this study investigated the score comparability of Vocabulary in Use test taken by 30 Iranian undergraduate students studying at a state university located in Chabahar region of Iran (CMU) to see whether scores from two administrations of testing mode were different. Therefore, two similar tests were administered to the male and female participants on two testing mode occasions with four weeks interval. Employing One-Way ANOVA statistical test to compare the mean scores and Pearson Correlation test to find the relationship between mode preference and performance revealed that two sets of scores were not different and gender difference was not also considered a variable that might affect performance on CBT. Based on the results, computerized version of the test can be considered a favorable alternative for the state undergraduate students in Iran.

Keywords: computer-based testing, paper-based testing, gender difference, test preference

1. Introduction

CBT has recently appeared as one of the most demanded viable form of alternative assessment throughout the world. Along with the development of computer assisted language learning (CALL) in education, applying computers as accepted assessment tools seems to be inevitable especially in academic settings. In education, CBT is used to evaluate the language proficiency of English learners (Fleming & Hiple, 2004). The IBM model 805 machine used in 1935 has been recorded as the first attempt to use computers in educational testing domain. It aimed to score objective multiple-choice item tests of American test takers each year to reduce the costs of scoring labor of millions of test takers throughout the USA. After publication of the first book on CBT in language domain (Holtzman, 1970), many developments in technology caused rapid enhancements in comprehensive language testing software packages to use great advantages of CBT such as the innovation, efficiency and productivity (Al-Amri, 2009). CBT assesses test taker's language proficiency accurately by providing more efficient standardization of test administration conditions (Al-Amri, 2009). In CBT, the same instructions, materials and information are presented in an enhanced consistent and uniform way to all test takers, regardless of the testing population size, place and time of testing. Moreover, unlike paper examination in conventional classroom, immediate viewing of scores on screen is provided in CBT session to give test takers the instant feedback. But, in some cases of large-scale CBT occasions, the security issues such as identity detection of test takers are the main concern. Wainer and Eignor (2000) have mentioned the security concerns even for computer-adaptive test or CAT in which the question items are adapted to the examinee's level and the items above or below the test takers' ability level are not submitted.

However, some institutions, testing organizations and universities have started to change the mode of testing administration and to replace their PPTs with CBTs in language assessment field (Kate, 2012). Empirical research on cross-mode comparability should be conducted to find out whether test scores across testing modes are equivalent in order to replace PPT with CBT. Although CBT offers some benefits over its traditional counterpart (Poggio, Glasnapp, Yang & Poggio, 2005) comparability and equivalency of test scores between the two test administration modes have been the real concerns for educators, scholars, practitioners and designers in assessment field (Lottridge, Nicewander, Schulz, & Mitzel, 2008).

Evaluating the comparability of PPT and CBT scores is a critical issue before introducing the computerized assessment into any educational context. The main objective of comparability study is to determine if test results obtained from two versions of the same test are equivalent. Mojarrad et al. (2013) conducted a comparability study on reading comprehension skill. They concluded that obtained scores across two various testing modes were not significantly different and test takers had positive attitudes towards onscreen version of the test. Furthermore, Boo et al. (2012) found that although test takers preferred computer counterpart of the conventional test, the scores received from CBT and PPT were comparable in terms of internal consistency, criterion and construct validities, means and standard deviations.

Choi, Kim, and Boo (2003) reported that the results of paper and computer versions of the standardized English language test administered to post-secondary level language learners were comparable. They proved that two versions of listening, reading comprehension, grammar and vocabulary subtests measured the same constructs based on confirmatory factor analysis results. Hosseini et al. (2014) conducted a comparability study and investigated the equivalency of test results obtained from CBT and PPT. Two equivalent multiple-choice tests of general English including computerized and paper-based formats were administered to one testing group composed of 106 Iranian English language learners who have been randomly selected from Azad University of Tehran. The findings of her research indicated that participants' performance on PPT was better than CBT performance.

Then, more attention should be paid to the testing mode effects on the equivalency of the scores that are obtained from two modes of presentation, i.e. traditional paper-and-pencil testing and computerized testing.

Several studies have been recently conducted to show that in order to replace computer-based test with its conventional paper-and-pencil counterpart, we need to prove that these two versions of test are comparable, in other words the validity and reliability of the computerized counterpart are not violated. In fact, the most critical problem that arises from converting PPT into CBT is validity. However, enough convincing evidence is not available to indicate that the CBT counterpart of a test may produce less valid results.

When comparing CBT and PPT, concerns also exist in regard to subgroups taking the tests. Gallagher, Bridgeman, and Cahalan (2002) examined data from testing programs such as GRE, SAT, Praxis, TOEFL, and GMAT, with regard to gender subgroup. They concluded that females performed better on PPT. Regarding to the subgroup testing, Wallace and his colleague declared that both genders outperformed on CBT (Wallace & Clariana, 2005, p. 176).

Students' preference may also be considered another critical factor in comparability studies. Due to the possibility of customizing the assessment based on personal preferences, some people prefer to take CBT version of the test. Although some students may prefer CBT, others may prefer PPT (Cater et al., 2010; Russell et al., 2010). It is a contributing factor that should be considered in comparability studies. Some test takers prefer PPT process because they are accustomed to taking notes and circling questions and/or answers for later review. In a research conducted by Flowers et al. (2011), there was a high preference for CBT, and test takers' preference had negative correlation with their performance on CBT. According to their findings, although test takers showed high preference for taking CBT, they outperformed on PPT. In a similar study, Higgins et al. (2005) reached to the conclusion that 87% of participants preferred to take CBT due to ease of use feature of CBT and no significant difference was also found between test takers' scores received from two versions of the test. The, based on their research results, no correlation was found between test mode preference and testing performance. According to another study done by Al-Amri (2009), although test takers preferred to take CBT, their test performance was better on PPT. His research findings showed no relationship between testing mode preference and test performance.

Indefinite findings about testing mode effects on validity and reliability of the test, and about the relationship of gender and testing mode preference with testing performance lead to the conclusion that using computers as preferred assessment tools to evaluate male and female language proficiency will continue to be debated. Computers are becoming more prevalent in individual daily life and making the life more automated. Increasing

use of computers in academic settings especially in language domain necessitates conducting more research on comparability and equivalency of scores received from two PPT and CBT modes of testing administration. Besides, the relationship of some external moderator variables such as gender and testing mode preference with testing performance should be considered with more attention. Therefore, the purpose of the present study is to find out whether test scores of selected Iranian undergraduate students are different across modes. It also attempts to investigate the relationship of gender and testing mode preference with testing performance. Thus, considering both theoretical and pedagogical perspectives, the following questions are addressed in this study to accomplish the main purposes:

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

RQ2: Is there any significant difference in test results of CBT between female and male ESP students in CMU?

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

2. Method

2.1 Participants

The population attended in the study was 30 people selected from the ESP students at Marine and Maritime University of Chabahar. This university is one of the two maritime universities and the biggest one in Iran and one of the biggest specialized universities in Middle East located in a major city on the South East coast of MAKRAN SEA (Gulf of Oman). The subjects were told that their responses to tests would be anonymous and the results would only be used for research purposes. New Interchange placement test was administered to 100 undergraduate students to discriminate intermediate level students. Based on the *common person design* which is a powerful design to collect good data for making score comparison and detecting differences in smaller sample of test takers, the 30 selected homogenous students were assigned to one group to take two versions of the same test with a four weeks interval. The age of participants ranged from 20 to 24 years old.

Table 1. Gender frequency distribution

Gender	Frequency	Percentage (%)
Male	24	80
Female	6	20
Total	30	100

Number of students participated in the study.

2.2 Instruments

The study employed intermediate level of Vocabulary in Use multiple-choice achievement test as the research data instrument to compare the scores received from both testing modes. The paper-based version of the English Vocabulary in Use test was converted into computer version using ClassMarker.com website. Another instrument to collect the research data concerning the third research question was a simple question mentioned at the bottom of test takers' exam paper and screen, i.e. *would you prefer taking the test on: paper – no difference – computer.*

2.3 Procedure

In the first testing session, the testing group was given the PPT version of the intermediate level of Vocabulary in Use multiple-choice achievement test. At the end of the exam, the testing group answered *would you prefer taking the test on paper – no difference – computer* simple question appeared at the bottom of the exam paper to explore the relationship between testing mode preference and testing performance. Then, the test takers' responses were scored. To minimize the practice and fatigue effects of testing, after a four weeks interval and in the second stage of the research, the testing group took the computerized version of the test. Before the exam, they received a simple sample computerized task and oral instruction on how to take the computerized version of the test. After becoming familiar with the CBT environment, every test taker was given a unique registration code to register into the assigned group created in the website. Test takers had 40 minutes to answer 50 question items (the time given to complete the sample exercise before administration of CBT was not included). On the onscreen test, students received one question per screen. Students clicked on the letter of the correct answer choice and then proceeded to the next question. Like paper-based testing, students could go back, review and

change previously answered questions in CBT. Like first stage, the testing group answered the simple question appeared at the bottom of the exam screen.

3. Results

Since internal consistency of research data collection instruments is important to say they are stable and consistent over time, a Cronbach's α reliability analysis was performed on the scores of test takers obtained from both PPT and CBT versions. The analysis indicated relatively high reliability coefficients (PPT, α=88 and CBT, α=87) (Table 2). Since our test might be affected by an increase in Cronbach's alpha, we looked for greater alpha values than our overall values by deleting particular items. In the analysis, the worst offenders were question item 4 for PPT and question items 10 and 32 for CBT version. They would increase the reliability values from .887 to .890 and from .877 to .879 for PPT and CBT, respectively that were not dramatic increases. Then researcher preferred to keep the content of test items. Both obtained values for reliability of the test reflected a reasonable degree of reliability.

Table 2. Internal consistency reliability

Testing Mode	*N of Questions*	*Cronbach's Alpha*	Alpha if Item Deleted	Item(s)
PPT	*50*	***.887***	***.890***	*4*
CBT	*50*	***.877***	***.879***	*10 &32*

Cronbach's coefficients of PPT & CBT.

Valid data samples that were collected from the testing group were all the 30 including 24 male and 6 female university students. By considering gender variable in this analysis, all participants' correct answers were counted and ported into version 22 SPSS for further analysis. To reach the goals of the present study, a quantitative approach including descriptive statistics and ANOVA test was used to answer the first research question by comparing the means of sets of scores. It was done to see if there was any difference between the scores of PPT and CBT. A majority of research conducted on PPT and CBT comparability study focused on the differences in means and standard deviations (e.g. Makiney, Rosen, & Davis, 2003; Pinsoneault, 1996).

Of the two versions of the test taken by the testing group, the highest mean score was found in PPT version, with a relatively higher mean score for PPT than for CBT by 0.53 points (Table 3). As displayed in Table 3, test takers' mean score on PPT (M = 46.66, SD = 17.43) was a little bit higher than their mean score on the CBT (M = 46.13, SD = 13.8). On the other hand, the standard deviation in PPT was higher than in CBT. It means that the dispersion of scores from mean score in PPT was higher than in CBT; consequently, it was concluded that Standard Error of Measurement (SEM) in CBT is lower than in PPT. Then, CBT had more consistent scores.

Then, one-way analysis of variance was conducted to examine the score differences with a null hypothesis of no difference. All the statistical analyses were done with a significant level of .05. Then, according to the findings of One-Way ANOVA test (Table 4), there was no statistically significant difference in scores from both PPT and CBT at a .05 level. Therefore, One-Way analysis of variance confirmed the null hypothesis that there was no statistically significant difference in the results of PPT and CBT versions of Vocabulary in Use achievement test administered to CMU ESP students. Based on the results of the score analysis of two testing sessions, the Sig. value was .896 at $P<0.05$. This amount of significance value at 29 (N-1) degree of freedom at a .05 level revealed no significant difference between two sets of scores obtained from two formats of the test and the test scores of participants are not different in paper-based and computer-based versions of the test (Sig=.896, $P>0.05$).

Table 3. Descriptive Statistics

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
PPT	30	**46.66**	**17.43**	3.18	40.1581	53.1753	26.00	98.00
CBT	30	**46.13**	**13.80**	2.52	40.9781	51.2885	20.00	64.00
Total	60	46.40	15.59	2.01	42.3723	50.4277	20.00	98.00

Distribution of participants' scores in PPT & CBT versions of the test.

Table 4. ANOVA Results

	Sum of Squares	D.F.	Mean Square	F	Sig.
Between Groups	4.267	1	4.267	.017	**.896**
Within Groups	14338.133	58	247.209		
Total	14342.400	59			

Comparison of test scores received from PPT & CBT versions.

Therefore, one way ANOVA analysis showed that the difference between the scores of PPT version of the test (n = 30, M = 46.66, SD = 17.43) and the scores of CBT version of the test (n = 30, M = 46.13, SD = 13.80) were not statistically significant, Sig = .896, p>0.05.

The second research question investigated whether CBT scores of female participants were different from the results of CBT scores of male participants. According to the distribution of male and female participants' scores in CBT (Table 5), the mean score of male participants on CBT (M = 45.66, SD = 14.98) was higher than the mean score of female participants on the CBT (M = 44.66, SD = 5.46). Of the two CBT mean scores, the highest mean score was found in male CBT, with a relatively higher mean score by 1 point. On the other hand, the standard deviation in male CBT was higher than in female CBT. It means that the dispersion of scores from mean score in male CBT was higher than in female CBT; consequently, it is concluded that Standard Error of Measurement (SEM) in female CBT is lower than in male CBT. Then, CBT had more consistent scores. According to the results of the analysis on male and female participants' scores in CBT, the Sig. observed value is .875 at P<0.05. This amount of significance value at 29 (N-1) degree of freedom in a .05 level revealed no significant difference between two sets of scores (Sig=.875, P>0.05) (Table 6).

Table 5. Descriptive Statistics

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Male CBT	24	**45.66**	**14.98**	3.05	39.3386	51.9947	20.00	64.00
Female CBT	6	**44.66**	**5.46**	2.23	38.9315	50.4019	38.00	50.00
Total	30	45.46	13.54	2.47	40.4094	50.5239	20.00	64.00

Distribution of male and female CBT scores.

Table 6. ANOVA Results

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

One-Way ANOVA comparing male and female CBT scores.

Therefore, one way ANOVA analysis showed that the differences between the male participants' scores in CBT version (n = 24, M = 45.66, SD = 14.98) and female participants scores in CBT version of the test (n = 6, M = 44.66, SD = 5.46) were not statistically significant, Sig= .875, p>0.05.

To investigate the relationship between test takers' testing mode preference and their testing performance, the correlation between participants' responses to the simple question appearing at the end of PPT exam and their CBT mean score was examined by Pearson product-moment correlation. The answers that participants gave to the question were coded as 1, 2 and 3 for "On paper", "No difference", and "On computer".

Table 7. Pearson Correlation

Pre-CBT testing mode preference	Pearson Correlation	-.153	Mean of CBT
	Sig. (2-tailed)	.419	
	N	30	

Correlation of pre-CBT testing mode preference and mean of CBT scores.

The results showed a negative and weak correlation between two variables. According to the results, test takers' testing mode preference (M=1.86, SD=.89) and their CBT testing performance (M=45.46, SD=13.54) were not strongly correlated (-.153(28) =.419, P>1 (Table 8). The strength of correlation as an effect size can be verbally described based on a guide (including "very weak" .00-.19, "weak" .20-.39, "moderate" .40-.59, "strong" .60-.79, "very strong" .80-1.0) describing absolute value of r that was suggested by Evans (1996). According to the findings, it can be concluded that changes in pre-CBT mode preference were weakly correlated with changes in test takers' scores in CBT.

Table 8. Pearson Correlation

Post-CBT testing mode preference	Pearson Correlation	-.176	Mean of CBT
	Sig. (2-tailed)	.352	
	N	30	

Correlation of post-CBT testing mode preference and mean of CBT scores.

The Pearson product-moment correlation was also run to examine the relationship between post-CBT testing mode preference and CBT testing performance. According to the results, for the testing group, the answers of participants to the second testing mode preference question (M=2.46, SD=.81) and their CBT testing performance (M=45.46, SD=13.54) were not significantly correlated, -.176(28) =.352, P>1 (Table 9). In the next stage, we examined whether participants performed better on their preferred testing mode based on their pre and post-CBT testing mode preferences and their relationship with testing performance.

Table 9. Descriptive statistics

Testing sessions	Preferred testing mode	N	Mean		Std. Deviation	
			Pre-CBT performance	Post-CBT performance	Pre-CBT	Post-CBT
PPT Ps	**Paper**	14	**46.85**	**47.71**	10.74	17.20
	No difference	6	**42.66**	**49.33**	6.77	10.93
	Onscreen	10	**55.60**	**45**	25.93	15.94
CBT Ps	**Paper**	6	**50**	**64**	12.52	49.33
	No difference	4	**48**	**48**	13.85	13.85
	Onscreen	20	**66**	**46**	16.87	16.87

To examine the relationship of pre-CBT testing mode preference of different preference groups with their testing performances.

According to the findings, before implementing CBT version of the test, PPT participants who preferred PPT version of the test (PPT performance, M=46.85) in the PPT session outperformed on CBT (M=47.71) and those who preferred CBT version of the test (PPT performance, (M=55.60 performed better on PPT (CBT performance, M=45). Consequently, those who did not mind taking the test on either mode, did better on CBT (M=49.33). After implementing CBT version of the test, the answers of testing mode preference questionnaire appeared at the bottom of the screen was also analyzed. Those CBT participants who preferred PPT version of the test (PPT performance, M=50) outperformed on CBT (M=64) and those who preferred CBT (PPT performance, M=46) performed better on PPT (M=66). The findings indicated that there was no positive interaction between testing mode preference of test takers and their testing performance. Then, it can be concluded that testing mode preference does not affect test validity. As the last step of this study, the influence of exposure to the CBT version of the test on participants' posterior testing mode preference was examined. To show the difference between testing mode preference before and after exposure to CBT, the answers of the participants to the testing mode preference question were summed up to show a frequency table of responses. In short, based on our findings, although test takers show high preference for taking CBT, they did better on PPT version of the test.

Table 10. Descriptive statistics

Preferred testing mode	(Pre-CBT) PPT		(Post-CBT) CBT	
	Frequency	Percentage	Frequency	Percentage
On paper	14	**46.66**	6	**20**
No difference	6	**20**	4	**13.33**
Onscreen	10	**33.33**	20	**66.66**
Total	30	**100**	30	**100**

Measuring differences between pre and post-CBT testing mode preferences.

From Table 10, 46.66%, 33.33% of participants preferred to take PPT and CBT versions of the test, respectively, before exposure to the CBT. Besides, 20% of participants didn't mind taking the test in either mode. After implementing CBT version of the test, only 20% still preferred to take PPT and 13.33% of the participants didn't mind taking the test in either mode. In this step of the study, the greatest percentage (66.66) was provided by the participants who chose CBT version of the test. The findings revealed that, after exposure to the CBT, the number of participants who preferred to take PPT and those participants who preferred to take the test in either mode changed in favor of the participants who preferred to take CBT.

5. Conclusion

The main objective of the present research was to find out whether there was statistically significant difference between the scores obtained from PPT and CBT versions of the same or equivalent test. In order to achieve this goal, two versions of Vocabulary in Use test were administered to the homogenous ESP students who were selected from the undergraduate students of Chabahar Maritime University, Iran. The received results and two sets of scores of test takers have been analyzed by SPSS statistical package to find out any difference between two testing modes. Test scores of test takers did not vary in both PPT and CBT. Like Hosseini et al. (2014) who confirmed that test takers received lower scores in CBT than PPT version of the achievement test, and Al-Amri who found cross mode effects (Al-Amri, 2008), findings of this study confirmed the comparability and equivalency of test takers' scores obtained from two different testing modes. The findings of this research are compatible with the corresponding findings that were reached by some other researchers (Mojarrad et al., 2013).

Another research question was to investigate the difference between testing performance of male and female participants on CBT. As the findings revealed, no significant difference was found for male and female test takers' scores across the modes. It supported the findings of Eid (2004) in which similar scores were obtained for male and female participants. His research was done with fifth grade participants who received similar scores in the math test implemented in two modes. Although the present research recommends that it is needed to conduct more studies on the gender issue in comparability studies, the findings of this study are a piece of evidence that confirm gender differences may not be a factor influencing testing performance on CBT.

The findings of this study indicated no correlation between testing mode preference and testing performance of test takers. The findings supported 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. Similar to the present study, no significant difference was found between test takers' scores on two versions of test which indicated no correlation between test mode preference and test performance (Higgins et al., 2005).

The findings of the present research are among the pioneers that have been conducted among undergraduate students of state universities in Iran. It has made a critical contribution to the literature by examining the effects of PPT and CBT versions of the same test on the achievement of participants from different cultures. Although computer technology has been widespread as a quick way of scoring and fast method of tracking the students' development in educational domain all over the world in 21st century, it seems that we are missing what we had in the past when teachers and parents devoted time to talk over their children needs. This point has been mentioned by Jamieson. He declared that rapid acceptance of technology developments may leaded to the belief that it would be all problem-solving advancement (Jamieson, 2005).

In this study, we also reached to the conclusion that although exposure to the CBT may change prior testing mode preference and may lead to the positive attitudes towards this kind of test version, the prior testing mode preference as an external moderator factor does not have influence on the CBT testing performance of the participants.

The present exploratory and experimental study aimed at showing the comparability between two versions of an English vocabulary test by assessing the potential values of certain statistical measures, indices and indicators applied in computerized model of English vocabulary knowledge test. Then, other actual computerized language skills assessment system may incorporate a more balanced representation of what is known about the acquisition of other language skills and sub-skills. Actually, the present limitations of computer analyses of human language do not make it possible to address directly the more important assessment of communicative competence.

Although some variables such as testing mode preference and gender were considered in this study, several others such as ethnicity, affective and motivational factors, test anxiety, test order effects, differences in testing conditions, cognitive processing, characteristics of computers being used, screen size and resolution, font characteristics, line length, number of lines, interline spacing, white space, scrolling, item review, and item presentation that may influence the measured performance of test takers in CBT version can be taken into account in future studies. It is worth mentioning that the results of the present research should not be generalized to the settings with more heterogeneous participants. Then, further replications of the study with more participants who are less homogeneous would be desirable thereafter. Finally, it should be mentioned that due to the limited number of female students and the higher percentage of male students studying in ESP courses at Chabahar Maritime University in comparison with female students, the researcher had to conduct the comparability study with those available homogenous ones. To gain more detailed and generalizable findings about the effect of gender difference variable in CBT testing performance, a larger female society is recommended for future studies.

Acknowledgments

This research was possible by supports of CMU that provided access to the university classes and computer laboratory. We would like to thank the members of CMU language department whose support, encouragement and expertise were great help. We thank with grateful recognition Dr. Nathan Thompson**,** Chief Product Officer at Assessment Systems Corporation whose expertise on assessment and educational technology assisted the research in every step of developing the work. We really appreciate his honestly guidance and informative comments.

References

Al-Amir, S., (2009). *Computer-based testing vs. paper-based testing: establishing the comparability of reading tests through the evolution of a new comparability model in a Saudi EFL context.* Unpublished doctoral dissertation. University of Essex, England.

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 & Linguistics, 10*, 22-44.

Boo, J. & Vispoel, W. (2012). Computer versus paper-and-pencil assessment of educational development: A comparison of psychometric features and examinee preferences. *Psychological Reports*, *111*, 443-460. https://doi.org/10.2466/10.03.11.PR0.111.5.443-460

Cater, K., Rose, D., Thille, C., & Shaffer, D. (2010, June). *Innovations in the classroom.* Presentation at the Council of Chief State School Officers (CCSSO) National Conference on Student Assessment, Detroit MI.

Choi, I. C., Kim, K. S., & Boo, J. (2003*)*. Comparability of a paper-based language test and a computer-based language test. *Language Testing,* 20, 295-320. https://doi.org/10.1191/0265532203lt258oa

Eid, G. K. (2004). An investigation into the effects and factors influencing computer-based online math problem-solving in primary schools. *Journal of Educational Technology Systems*, *33*(3), 223-240. https://doi.org/10.2190/J3Q5-BAA5-2L62-AEY3

Fleming, S., & Hiple, D. (2004). Foreign language distance education at the University of Hawai'i. In C. A. Spreen, (Ed.), *new technologies and language learning: issues and options* (Tech. Rep. No.25) (pp. 13-54). Honolulu, HI: University of Hawai'i, Second Language Teaching & Curriculum Center.

Flowers, C., Do-Hong, K., Lewis, P., & Davis, V. C. (2011). A comparison of computer-based testing and pencil-and-paper testing for students with a read- aloud accommodation. *Journal of Special Education Technology*, *26*(1), 1-12. https://doi.org/10.1177/016264341102600102

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), 133-147. https://doi.org/10.1111/j.1745-3984.2002.tb01139.x

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).

Holtzman, W. H. (1970). Individually tailored testing: Discussion. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance*. New York: Harper & Row.

Hosseini, M., Zainol Abidin, M. J., Baghdarnia, M., (2014). Comparability of Test Results of Computer-Based Tests (CBT) and Paper and Pencil Tests (PPT) among English Language Learners in Iran. *International Conference on Current Trends in ELT*, 659-667. https://doi.org/10.1016/j.sbspro.2014.03.465

Jamieson J. M., (2005). *Annual Review of Applied Linguistics, 25*, 228-242. https://doi.org/10.1017/S0267190505000127

Kate Tzu, C. C. (2012). Elementary EFL teachers' computer phobia and computer self-efficacy in Taiwan. *TOJET: The Turkish Online Journal of Educational Technology*, *11*(2).

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.

Makiney, J. D., Rosen, C., Davis, B. W., Tinios, K., & Young, P. (2003). *Examining the measurement equivalence of paper and computerized job analyses scales*. Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Mojarrad, H, Hemmati, F, Jafari Gohar, M., & Sadeghi, A. (2013). Computer-based assessment (CBA) vs. Paper/pencil-based assessment (PPBA): An investigation into the performance and attitude of Iranian EFL learners' reading comprehension. *International Journal of Language Learning and Applied Linguistics World*, *4*(4), 418-428.

Pinsoneault, T. B. (1996). Equivalency of computer-assisted and paper-and-pencil administered versions of the Minnesota Multiphasic Personality Inventory-2. *Computers in Human Behavior*, *12*, 291-300. https://doi.org/10.1016/0747-5632(96)00008-8

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.

Russell, M., Almond, P., Higgins, J., Clarke-Midura, J., Johnstone, C., Bechard, S., & Fedorchak, G. (2010,). *Technology enabled assessments: Examining the potential for universal access and better measurement in achievement.* Presentation at the Council of Chief State School Officers (CCSSO) National Conference on Student Assessment, Detroit MN.

Wainer, H. & Eignor, D. (2000). *Caveats, pitfalls and unexpected consequences of implementing large-scale computerized testing. Computer adaptive testing: A primer* (2nd ed, pp. 271-99). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wallace, P., & Clariana, R. (2005). Perception versus reality–Determining business students' computer literacy skills and need for instruction in information concepts and technology. *Journal of Information Technology Education, 4*, 141-151.