Comparability of Test Results of Computer Based Tests (CBT) and Paper and Pencil Tests (PPT) among English Language Learners in Iran

Monirosadat Hosseini\textsuperscript{a}, Mohamad Jafre Zainol Abidin\textsuperscript{b}, Mostafa Baghdarnia\textsuperscript{c,*}

\textsuperscript{a} \textsuperscript{b} \textsuperscript{c} University Sains Malaysia (USM), 11800, Penang, Malaysia

Abstract

This study aims at examining the score comparability of institutional multiple-choice reading comprehension tests in two testing methods, i.e. paper-based and computer-based tests taken by Iranian first-year English students in Azad University of Tehran, Iran. In order to find the results, the researcher required examining the impact of computer-based testing (henceforth CBT) on the test score results, and exploring the relationship between particular test takers’ characteristics such as prior computer familiarity and computer attitudes as well as test performance with their test scores. Two equivalent tests were administered to participants on two different occasions. Utilizing matched t-test to compare the means of two test modes, the results of the study show the priority of PPT over CBT with .01 degree of difference at p<05. Using ANOVA, the findings revealed that computer familiarity and attitude towards computer had no significant influence on the students’ performance in computerized test. Additionally, participants showed more preference on test features presented on the computer test.

Published by Elsevier Ltd. Open access under CC BY-NC-ND license.

Selection and peer-review under responsibility of Urmia University, Iran.

Keywords: Computer-based test (CBT); Paper-based test (PBT); computer familiarity; computer attitude; test performance, test preference

1. Introduction

This study aims at examining the score comparability of multiple-choice General English achievement tests in two computer-based testing (henceforth CBT), on the test performance of students, and find out the relationship between

* Corresponding author. Tel.: +989192477783; fax: +982122442041.
E-mail address: ho.mahmonir@yahoo.com
different testing modes, i.e. paper-based tests and computer-based tests taken by Iranian English Translation students in Azad University of Tehran – Iran. It also aims at examining the impact of the new testing mode, i.e. some specific test takers’ characteristics such as computer familiarity, attitudes towards computers, and test preference of students. The results could be applied in other universities and language institutions.

2. Computer Based Tests versus Paper and Pencil Tests

With the appearance of new technologies, computerized testing has begun to be widespread and implemented in large scale testing. However, the limited accessibility of computer and high cost limited the implementation of computerized language testing. Along with the development of computer assisted language learning (CALL), the need for applying computer in language testing (CALT) is inevitable in academic contexts. Especially, since the language proficiency standard tests such as TOEFL and IELTS have been changed into computerized formats, students need to be familiar with such occasions before encountering such proficiency test modes.

Moreover, in language learning, the use of computers and electronic devices has become popular all around the world, especially in evaluating the language proficiency of English learners, the most precise and available way is through computers and on-line process (Fleming & Hiple, 2004). Such developments in computer technologies have influenced many areas including educational settings such as online learning, testing and assessment (Bennett, 2002; Pommerich, 2004).

Some students after taking such computerized tests complain that their test score is not the real representative of their language proficiency because of their unfamiliarity with such test modes. Despite the use of computer in language learning, the examinations are still conducted in the traditional form, i.e. paper based format in most academic contexts including Azad University in Iran.

However, as institutions started to accomplish computer-based testing in their examination systems along with traditionally paper-based testing systems, concerns arose about the comparability of scores from the two administration modes (Wang, 2004). As the computerized tests have been used for almost 20 years (Laborda, 2007), and the computer assisted language learning (CALL) has been common since the middle of 20th century, it has been necessary to develop the means to include computerized tests (Leahy, Lyon, Thompson, & William, 2005). Russell and Haney (2000) asserted that the “mismatch between the mode of learning and assessment could cause achievement to be inaccurately estimated.” (p. 2).

Although CBT offers many advantages over traditional PPT (Poggio, Glasnapp, Yang and & Poggio, 2005; Zhang & Lau, 2006), equivalency of scores between the two test administration modes have been the real concerns for the researchers and experts in the area of assessment, practitioners, and educators (Lottridge, Nicewander, Schulz, & Mitzel, 2008). To deal with this concern, many researchers conducted studies in synthesizing the administration mode effects of CBTs and PPTs (Pommerich, 2004; Poggio et al., 2005; Salimi, Rashidy, Salimi, & AminiFarsani, 2011).

Noyes and Garland (2008) believe that the benefits of standardized computer-based tests, such as quick and objective results and the ease of comparing results with others make this method very popular. Moreover, moves toward computerized testing are rooted in the advantages it provides in comparison with traditional paper-and-pencil format (Choi & Tinkler, 2002; Kim & Huynh, 2007; Kingston, 2009). Such advantages, according to the findings of mentioned studies, include cost-effective administration, ease of administration, more accuracy, immediacy of scoring and reporting, and flexible test scheduling and location. These studies, also, indicated that students who are familiar with computers feel more comfortable using it (O’Malley, Kirkpatrick, Sherwood, Burdick, Hsieh, & Sanford, 2005; Poggio, et al., 2005).

While computers have been important in language testing, only a relatively small group of professional language testers use computers in producing and administering language tests (Kathleen, 2006). However, scores
derived from CBT as compared to PPT might reflect not only the examinee’s proficiency in the construct being measured, but also the level of language proficiency of examinees (Clariana & Wallace, 2002). Clariana and Wallace (2002) in their study found that: “It is critical to realize that computer-based and paper-based tests, even with identical items, will not necessarily produce equivalent measures of student learning. Instructors and institutions should spend the time, cost, and effort to mitigate test mode effects.” (p.44)

1.1. Key Factors Associated with Test Mode Effects

Fulcher (2009, p.7) states that, "Validity is conceptualized as test recognition by institutions, and validity evidence becomes the extent to which linkage is demonstrated by institutionally approved procedures" (North, Figueras, Takala, Van Avermaet, Verhelst, 2003).

The problem with computerized tests arises when the matter of validity comes; however, there is no evidence to show that the construct of CBT may produce less valid tests. Instead, other factors may influence tests that have little to do with the testing objectives the test developer intends to provide. For example, in many computer-based tests, it seems that the test designer started from a valid objective, but the limitations of the program, system, language or the tester's own characteristics have influenced the results of tests (Chapelle & Douglas, 2006).

As pointed out by other authors in the special issue of using technology in education, for many reasons, the use of computer-based assessment is becoming widespread (Poggio, 2005; Lottridge. 2008; Yurdabakan, 2012). The examples of conducting exams through computers include entrance exams in postsecondary education, state drivers’ license tests, examination of military training, job application exams in the private sector, and certification exams by professional groups (Russo, 2000; Trotter, 2001). However, the results of many studies in the comparability of PPT and CBT show that there is no empirical evidence that identical paper-based and computer-based tests produce the same results. Clarion and Wallace (2002) suggested that the factors that may influence the test results rather than the construct being measured are known as the "test mode effect".

For example, paper-based test scores were greater than computer-based test scores for both mathematics and English tests in Mazzeo and his colleagues’ (1991) study. While computer-based test scores were greater than paper-based test scores for a dental hygiene course unit in midterm examination (De Angelis, 2000), some studies have reported non-significant difference between computer and paper-based test results (Schaeffer Reese, Steffen, McKinley, & Mills, 1993; Mason, Patry, & Bernstein, 2001). In regard to such different results in comparability studies of PPT and CBT, Yurdabakan (2012) believes that even though computer access opportunities increase students’ computer competencies and CBT achievements (Pomplun & Custer, 2005; Pomplun et al. 2006; Bennett et al. 2008), it is possible to evaluate the influence of students’ limited access opportunities on examinees' test performance. Similarly, Terzis and Economides (2011) stress that, besides many other factors, attitudes towards CBT applications are under influence of social environment and educational systems. On the other hand, attitude towards using computer and technology in educational settings may depend on educational contexts.

Leeson (2006) identifies the factors leading to difficulties in CBT applications under two titles, as factors originating from “users” and “technology used”. He states that the user’s gender, the ability to process information, the ability to use a computer, and the level of anxiety could have an influence on an application. He gave the size and resolution of monitors, writing character and its length, the way the problem is presented, and having the option of review or not as technology originated factors.

Many researchers have already done studies investigating the relationship between computer usage ability and achievement tests. Yurdabakan (2012) identifies some of these studies (Goldenburg & Pedulla, 2002; Pomplun & Custer, 2005; Pomplun et al., 2006; Bennett et al., 2008) stressing that computer usage ability is an important predictor of respondent achievement; therefore, those poor students at computers may show low achievement in CBT. However, they added that with the increase in computer technologies and access opportunities, such problems might decrease.
In some of the studies investigating the difference between methods in terms of gender, race and age, such as Clariana & Wallace, 2002 and Bennett et al., 2008, no significant difference was found in achievements, whereas in some other studies (Gallagher, Bridgeman, & Cahalan, 2002) little significant difference was observed. In their recent study, Terzis and Economides (2011) describe the trends of male and female students towards CBT. Yurdabakan (2011) reported that as a result, they highlighted both genders having positive views on CBTs, if they are based on games, they are open and course-related. Additionally, they state that the attitudes of males and females towards CBTs are under the influence of social environment; the boys focus on the usefulness while the girls stress the ease of use.

Boo (1997), in his study on the comparability of PPT and CBT, found that there was no relationship between computer familiarity and test performance of students in three computerized reading tests. Taylor and his colleagues (1999) also after examining the relationship between computer familiarity and test performance of 1,169 participants from different countries on TOEFL CBT, found no relationship between computer familiarity and examinees' test performance on TOEFL CBT.

due to the increasing use of computer in academic contexts especially in language learning, there have been many investigations on the comparability of test scores in two different test modes some of which consider different key factors influencing test results such as computer familiarity, prior attitude towards using computers, age, gender, and some other factors. The aforementioned studies on the comparability of learners’ performance on PBTs and CBTs on the one hand, and the fact that the current study is the first large scale research on CBT in universities of Iran on the other hand, have motivated the researchers to carry out this study. Therefore, the present study attempts to find out whether there is any relationship between key factors of computer familiarity and attitude towards computer and test scores in CBT among Iranian English students in Azad University of Tehran, Iran. The following research questions have been developed accordingly:

RQ1: Is there any statistically significant difference between Iranian English learners’ performance on computer-based tests and their paper-based ones?

RQ2: Is there any relationship between Iranian English students' computer familiarity and their performance on the computer-based test results?

RQ3: Is there any relationship between Iranian English students' attitudes towards CBT and their test performance in CBT?

3. Methodology

1.2. Participants and Instruments

The participants were 106 Iranian English students having been selected randomly from Azad University of Tehran. Instructors agreed to cooperate and obtain the agreement of their students to participate in the study. To ensure the homogeneity of participants, they were selected from second year students. One group was given two equivalent multiple-choice tests derived from General English Book in two different occasions, one in paper based type and the other in computerized format. It is noteworthy that a questionnaire derived from Computer Attitude Scale (CAS) developed by Loyd and Gressard (1984) and validated by Berberoglu and Calikoglu (1992), was distributed among the participants to get the information about their familiarity and attitude towards computer tests. The first part of the questionnaire elicited demographic information of participants including age and gender. The second section of the questionnaire consisted of 30 items to gauge the learners’ familiarity and attitude towards
1.3. procedure

In the first stage of the study, the paper-based format of achievement test was given to the respondents as the usual standard mode. The allocated time to answer the questions was 60 minutes. In the next stage, the computerized equivalent test was administered after two weeks. Before conducting the exam, the respondents were given some instruction about how to answer the computerized questions. The time they were given to answer the computerized achievement test was 60 minutes too. Thereupon, they were asked to fill out the questionnaire. It is worth mentioning that two equivalent tests had been piloted before administering the exams. In addition, to confirm the questionnaire data, 10 participants were randomly selected among the total number of 106 participants who filled out the questionnaire and agreed to be interviewed. The interviews were conducted in Persian individually after the questionnaire data were collected. To keep the interview manageable, its length was limited to 5-10 minutes. The interview was about their preference in testing mode and the reason of changing their preference and its relationship with their test performance in both PPT and CBT.

4. Results and Discussions

To answer the first question of the study, at the beginning, descriptive statistics analysis was used to gain a better view of the data, and then the inferential statistics analysis was run to find out the relationship between mean scores.

Table 1 shows descriptive statistics of the study.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>N</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPT</td>
<td>24.16</td>
<td>106</td>
<td>4.5376</td>
<td>.44036</td>
</tr>
<tr>
<td>CBT</td>
<td>23.16</td>
<td>106</td>
<td>5.85074</td>
<td>.56827</td>
</tr>
</tbody>
</table>

As displayed in Table 1, the students' mean score on PPT (24.16) is higher than their mean score on the CBT(23.16). On the other hand, the standard deviation in PPT is lower than CBT. It means that the dispersion of the scores from mean score in CBT is higher than PPT; consequently, Standard Error of Measurement (SEM) in PPT is lower than CBT. The results of the descriptive statistics are shown in figure 1.

Since two means in the study came from the same subjects, matched t-test formula was run to compare the mean scores of the students on both tests. As can be seen from Table 2, the t-observed value is 1.99 at P < 0.05. This amount of t-value at 105 (N-1) degrees of freedom in a<.05 is greater than the critical value of $t$, i.e. 1.98.

The results of inferential analysis showed a significant difference between the students' mean scores on paper-based and computer-based tests.

Based on these results, it can be concluded that there is a statistically significant difference between the mean scores of the students on PPT and CBT.

In order to answer the second and third questions, ANOVA has been run to find out the the significance of difference between computer familiarity and attitudes towards computer and test performance of students. As for the second question of the study, the results in Table 3 show that the F Observed value for the students’ prior familiarity with computers and computer-based tests is 1.92 ($P =0.14 > 0.05$). Based on these results, it can be concluded that the students’ computer familiarity does not have any significant effect on their performance in their achievement test.
Previous results asserted that there is neither relationship nor interactive effect of prior computer familiarity of participants with its two measures on participants’ performance on computer-based tests. This implies that whether the subject has a high or low degree of computer familiarity, there is not any advantage or disadvantage while performing on computer-based tests. Furthermore, this also supports the construct validity of the computer-based tests as this construct-irrelevant variable is not part of the construct measured by the computer-based tests. It is worth mentioning that all participants were given a brief tutorial about CBT before getting involved in the real testing sessions to minimize the computer familiarity effect, overall, this answers the second question.

The F observed value for the effect of the students' attitudes towards computer on their performance in paper and computer based tests is 1.77 ($P = .17 > 0.05$). Based on these results, it can also be concluded that the students' attitudes towards computer does not have any significant effect on English students’ test performance. Based on these statistics, there is not any significant correlation or interactive effect between the participants’ prior computer attitudes and their performance on computer-based tests. This result showed no significant difference between students' attitude towards computer and their test performance in CBT and PPT; and hence positively answers the
second research question.

Table 3: ANOVA results of interactive effect of computer familiarity and attitude on computer-based test scores

<table>
<thead>
<tr>
<th>Source</th>
<th>Source</th>
<th>Df</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>mode * computer familiarity scale</td>
<td>Sphericity Assumed</td>
<td>1</td>
<td>1.92</td>
<td>.995</td>
</tr>
<tr>
<td>mode * computer attitudes scale</td>
<td>Sphericity Assumed</td>
<td>1</td>
<td>1.97</td>
<td>.326</td>
</tr>
</tbody>
</table>

In order to support the findings of the study, the researcher conducted an interview with 10 volunteer participants after administering both tests. The researchers asked about their preference of mode delivery tests. The results showed that almost 53% of the participants preferred CBT to PPT in spite the fact that the mean score of PPT was higher than CBT. Three participants justified their paper test preference by having no prior CBT experience. Two participants also attributed their preference of PPT to past unpleasant experiences such as boring computer courses or uncustomary of CBT in academic contexts. Furthermore, being accustomed to paper tests, the novelty of CBT was another reason for three other participants. Eye fatigue was a major concern for one participant from his prior experience with the daily use of computers. Those who preferred CBT to PPT found CBT more comfortable, more enjoyable, and time saving. Ease of changing answers, reading the passages and questions, as well as being able to navigate through the text and the questions were very attractive features of CBT that influenced the participants to change their testing mode preference. Participants also liked the display of the passages and the questions which was an innovation for them in the test-taking experience, as well as a shift from the classical testing mode (PPT) to the new technological one (CBT).

5. Conclusion

The results of the study revealed that participants performed better on paper-based tests than computer-based test (MPPT=24.6 > MCBT=23.6). The findings are in contrast with those who argue that there is not any difference between CBT and PPT if the test administration condition is equivalent except the influence of students' preference in computer-based test rather than paper-based test mode (Bachman, 2000; Jamieson, 2005; Chapelle, 2007; Douglas & Hegelheimer, 2007). However, it supports the previous findings that favor the students' better performance in paper-based test in comparison with computer-based tests (Coniam, 2006; Cumming, et al., 2006; Salimi et al. 2011). Therefore, this inconsistency in administration of the different modes of test provide support for the claim mentioned in the literature as there are controversies among the scholars in the field concerning whether test takers preferred computer-based over traditional paper-based tests (Sawaki, 1999; Bachman, 2000; Wolfe & Manalo, 2005). Therefore, the researchers would like to include both test administration formats in the language examination in universities. However, the results provide justification for including computer-based tests in traditional test administration. Moreover, computer technology has continued to be widespread into the 21st century as a crucial and versatile instrument for communication and education. However, rapid technological advancement can create a tendency towards acceptance of innovation and the belief that technology will be useful and solve all problems (Jamieson, 2005). This view can create problems, particularly if educators fail to act and react to the needs of learners. Thus, students who do not have positive attitudes towards learning English through computers will be at a distinct disadvantage.
Based on the results of the current study, the Iranian English students showed positive attitudes towards computer-based tests while they performed better on the paper-and-pencil based test. The most problematic issue here is the contrast between the preference of participants towards computerized tests and their better performance in PPT. The result implies the need of further research in this area considering factors that may influence the test performance of students in CBT and PPT. The findings may direct the instructors and test developers to improve the conditions of conducting CBT in academic contexts in Iran including universities and language institutes.

References

Bachman, L. F., (2000). Modern language testing at the turn of the century: Assuring that what we count counts. Language testing, 17(1), 1-42.


