Computer-based vs. Paper-based Testing: Does the test administration mode matter?

Saad Al-Amri
University of Essex, UK
salamr@essex.ac.uk

Abstract

Evaluating the comparability of paper-based and computer-based tests is crucial before introducing computer aided assessment into any context. There have been several comparability studies that have examined the impact of transferring a test from paper to screen. Such studies have either focused on the comparability of the product of the tests i.e. scores, or on the processes used to achieve that product. However, Chapelle & Douglas (2006:46) argue that "To date not a single study of L2 testing has examined directly whether or not past experience with computers affects test performance on a computer-based L2 test". Sawaki (2001) recommended that this type of empirical work should utilize different methodologies such as eye movement, verbal protocols, post hoc interviews and questionnaires in order to obtain useful results. Chalhoub-Deville and Deville (1999) pointed out that there is a scarcity of comparability research on localised language tests needed to detect any potential impact of the test delivery mode when converting conventional paper tests to computerised tests. Thus, our ongoing study explores the comparability of paper and computer-based testing in an L2 reading context and the impact of test takers' characteristics, i.e., computer familiarity, computer attitude, testing mode preference and test taking strategies on students' performance on computer-based tests, and in comparison with paper-based tests. 167 Saudi medical students participated in this study. The study used several quantitative and qualitative instruments to gather data. This paper reports on the results of the quantitative instruments of the study i.e. the tests and the questionnaires and the relevant interviews. We found no significant difference between each testing mode and none of the factors examined had an influence on students' performance when doing the computer-based tests. The study is still in the process of analyzing the qualitative data and hopes to report on that soon.

Introduction and literature review

Technology has been implemented in the field of language assessment by using computers to deliver different types of assessment. However, little empirical work has been done in order to examine the impact of technology on the main basic quality concepts in the assessment, which include the concepts of validity and reliability. Moreover, little research has been conducted to investigate the interaction between the assessment modes and test taker variables. There have been some studies that have focused on the comparability of paper-based testing and computer-based testing in some areas such as psychology, mathematics and ergonomics (Sawaki, 2001). However, only a few studies have explored this issue in the field of language assessment, such as those by Boo (1997); Taylor et al. (1999); Kirsch et al. (1998); Taylor et al. (1998); Eignor et al. (1998); Russell (1999); Russell and Haney (1997) and Choi et al.. (2003). Some studies have revealed that there is a significant difference between the two testing modes (Pomplun, 2002; Choi, et al., 2003) while others have concluded the opposite (Boo, 1997; Whitworth, 2001; Bugbee, 1996). However, previous research has mainly focused on the product, i.e. test scores achieved or the processes that resulted in these scores, but not on both aspects. Paek (2005) mentioned the importance of computer familiarity and test taking strategies on measuring the equivalence of the two testing modes.

The advantages of computers are well-known and apparent. They offer test developers the opportunity to improve their productivity and lead to innovation in their fields. The advantages of computers in the context of this research, i.e. language testing, are myriad. The standardization of test administration conditions is one of the benefits offered by computer-based testing (CBT). No matter what the tests' population size is, CBT helps test developers to set the same test conditions for all participants. It also improves all aspects of test security by storing questions and responses in encrypted databases and enables testers to create randomized questions and answers from vast question pools. Moreover, offering different test formats and the immediate presentation of different types of feedback, either to students or testers, are also some of the great advantages of CBT.

Collecting different performance data such as latency information is a unique feature of CBT (Olsen et al., 1989). On the examinees' side, they are able to receive greater measurement efficiency and the possibility to take the test at any time. On the other hand, there are some disadvantages that users have to be aware of before opting for computer-based testing. Students need some degree of computer literacy in order to avoid the mode effect on computer-based testing (Alderson, 2000). Johnson & Green (2004:2) asserted that "If computer technology is to be able to fulfil the potential claimed by its supporters, it needs to be seen to at least match the levels of validity and reliability of the paper and pencil assessments that it hopes to replace". Thus, many scholars suggest conducting systematic studies to check equivalency and comparability of paper-based tests and computer-based tests (Parshall et al., 2002).

One of the main contributing factors that should be examined when dealing with comparability research is the existing computer familiarity of test takers and its interaction with performance on CBT. Little research has been carried out in the area of the relationship between the computer familiarity of examinees and their performance on computer-based testing. Furthermore, the concept of computer familiarity has been defined in different ways (Taylor et al., 1999). In an extensive review of the relevant literature, Taylor et al. (ibid) found that the concept of computer familiarity has encompassed computer use (Pelgrum et al., 1993), computer experience (Geissler & Horridge, 1993; Hicks, 1989; Jegede & Okebukola, 1992; Levin & Gordon, 1989; Loyd & Gressard, 1984a; Marcoulides, 1988; Miller & Varma, 1994; Powers & O'Neill, 1993), awareness of technology and information technology (Christmas, 1992; Dalton, 1994; Durndell & Lightbody, 1993; Jegede & Okebukola, 1992) and having access to computers at home, school or elsewhere (Durndell & Lightbody, 1993; Geissler & Horridge, 1993; Levin & Gordon, 1989; Miller & Varma, 1994; Okinaka, 1992; Stephens & Rowland, 1993). Some researchers have found that computer familiarity can affect the examinees' performance on CBT (Buderson et al., 1989; Hofer and Green, 1985 and Mazzeo & Harvey, 1988).

It has been found that computer experience was a major factor in explaining the difference between students' performance on computer-based arithmetic reasoning tests (Lee, 1986). However, Boo (1997) found that there was no significant relationship between computer familiarity and the students' performance on three computerised tests. Moreover, Taylor et al. (1999) found no evidence of an undesirable effect of computer familiarity on students' performance on computer-based tests. Due to a high exposure to technology and the availability of computers, measuring computer familiarity has been a difficult issue in all of the previous research (Boo, 1997).

Furthermore, an essential additional test taker characteristic that might affect his/her performance with regard to CBT is computer attitude. There is no doubt that the examinee is highly influenced by his/her attitude or preference towards the test or the test mode. Some people are not familiar with technology and cannot keep pace with its rapid development and thus they prefer not to tackle or deal with any form of technology nor apply it in their academic or social lives. Computer attitude and preferences are not only formed by previous experience and use of computers but also by the educational and professional curricula and generally by choices and attitudes to subjects in schools (Bear, Richard and Lancaster, 1987, cited in Levin and Gordon, (1989)).

Different studies have explored examinees' computer attitudes and preferences for computer-based testing and found a variety of views. Some participants negatively evaluated their experience with the computers in general and CBT in particular (Ward et al., 1989). However, that was explained by the investigators as the respondents were new to this type of test administration mode and such a negative attitude might disappear with more exposure to CBT. On the other hand, many other studies found that the examinees positively preferred CBT for several reasons such as time efficiency, focussing attention, enjoyment and confidentiality (Bresolin, 1984, cited in Boo, (1997)). Other participants were very positive about computer-based testing because it seemed less difficult, more useful and engaged their attention more than paper-based testing (Harrel et al., 1987). Further studies concluded that students positively preferred computerised tests to their counterparts on paper and some studies related that to computer experience which means the greater the computer experience of the examinee, the more positive the attitude and the preference, whereas others changed their attitudes after exposure to CBT (Vincino and Moreno, 1988; Levin & Gordon, 1989; Bruke et al., 1987; Powers and O'Neill, 1992 and Boo, 1997).

Therefore, this study aims to measure the comparability of computer-based and paper- based tests, and the relationship with the two core concepts of assessment i.e. validity and reliability. This study also examines how test taker characteristics such as computer familiarity, computer attitude, testing mode preference, and test taking strategies interact with the testing mode, and to what extent this interaction affects the test scores and, as a result, the overall validity of computer-based tests. The methodology used in this study differs from previous research as the framework employed here is both quantitative and qualitative in nature. This framework triangulates the data sources to increase the validity and reliability of the results and the conclusions of this study. The conclusions and recommendations will be beneficial to medical colleges in Saudi Arabia as they are the main audience in the target context.

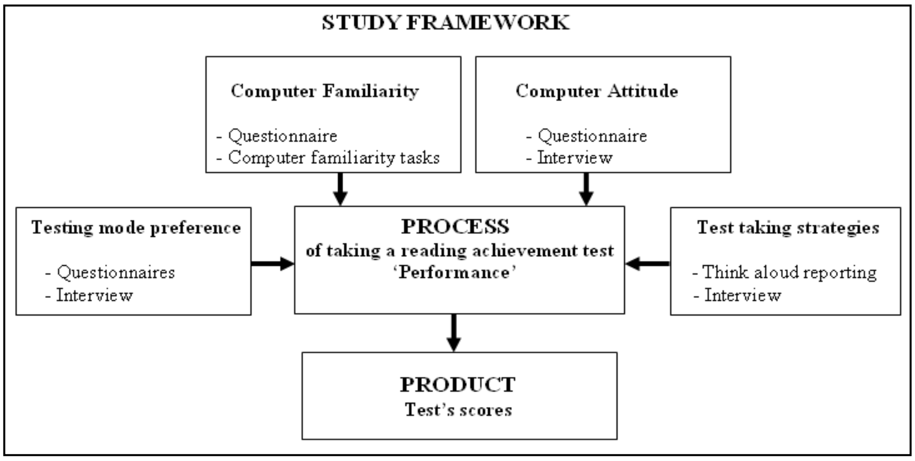
Research questions

This study addresses and attempts to answer the following questions:

1. Are the reliability and validity of the tests influenced by the test administration mode?
2. To what extent does prior computer familiarity affect participants' performance on CBT?
3. To what extent does prior computer attitude affect participants' performance on CBT?
4. Will participants' prior testing mode preference influence their performance on both testing modes? If so,
5. Will participants perform better on their preferred test mode?
6. Will participants' posterior testing mode preference be influenced by exposure to CBT?
7. Will CBT experience result in a positive attitude towards features of CBT?

Methodology

This study used a triangulation perspective in data collection from students doing both paper- and computer-based reading tests in order to vary the sources of data required to answer the research questions. Thus, both quantitative and qualitative instruments were employed to reveal more valid and reliable results and consequently reach more solid conclusions. The following diagram illustrates the framework used in this study:



Subjects

The participants in this study were first year medical students. 167 Saudi male and female students took part in this study. 57% of them were male and 43% were female. Participants had the same educational history with respect to reading instruction and test taking. Yet, female students had not received formal computer instruction in their academic curricula. This project was carried out at King Faisal University, College of Medicine, Dammam, Saudi Arabia.

Study instruments

The study employed the following instruments:

- Reading section of the TOEFL test (with permission from ETS).
- Two computer familiarity tasks designed by the researcher to measure, in a direct way, the participants' computer familiarity.
- Two questionnaires measuring computer familiarity, computer attitude, testing mode preference and attitudes to CBT features designed by the researcher with the help of questionnaires available in the literature.
- Three institutional reading achievement tests made available to the researcher by the target institution.
- Semi-structured interviews conducted after administering the second questionnaire and the verbal reports.
- Think aloud protocols to gain insights into mental processes used by subjects while taking the tests.

Procedure

Initially the reading section of the TOEFL test was administered to measure the students' reading proficiency and to measure the concurrent validity of the study tests. Then the computer familiarity tasks were given to the students to measure, in a more direct way, their computer familiarity and to validate that with their responses to the first questionnaire. That was followed by the first questionnaire which was designed to collect demographic data, measure the computer familiarity of participants, and their preference with regard to testing modes. Next, the study utilized three institutional achievement tests as a tool to compare the scores on both testing modes. These tests were converted into computer versions using the Question Mark system *Perception*. There were two testing sessions where all paper- and computer-based tests were administered in a counterbalanced design in order to minimize the practice and order effect with a one-month gap to minimize the memory effect as well. The second questionnaire was administered after that to collect data about preferences after exposure to computer-based testing. That was followed by semi-structured interviews with a random sample from the participants.

Results and discussion

The data have been collected recently and the coding and analyses have not yet been completed. However, initial analyses have been conducted for the purpose of this paper. First, the test scores analysis showed that there is a difference in students' performance on paper- and computer-based tests. Table 1 shows a summary of all tests' mean scores and standard deviations.

Table 1: Descriptive statistics of the paper and computer-based tests

Tests	N	Mean %	Std. Dev'n
Paper-based test 1	167	77.45	12.879
Paper-based test 2	167	66.47	17.856
Paper-based test 3	167	67.90	18.362
Computer-based test 1	167	74.65	15.389
Computer-based test 2	167	61.89	19.022
Computer-based test 3	167	64.07	15.259

Based on Table 1, we can see that the mean of every paper test is close to its computer counterpart. However, the t-test analysis proved that these differences are significant. Table 2 summarises the t-test results:

Table 2: T-tests of means of all paper and computer tests

		Sig.
Pair 1	paper-based test 1 - computer-based test 1	.016
Pair 2	paper-based test 2 - computer-based test 2	.001
Pair 3	paper-based test 3 - computer-based test 3	.003

We would argue that these significant differences are attributed to the low number of test items on each test and to the sample size. Moreover, it is quite obvious from the descriptive statistics Table of the tests that the difference between the means of each test is not vast. To examine the effect of the testing mode on reliability, we examined the internal consistency

(Cronbach's alpha) of each test on each mode. The following Table summarizes the results of the internal reliability coefficients:

Table 3: Reliability coefficients of all paper and computer tests

Tests	Testing Mode	
	Paper	Computer
Test 1	0.57	0.58
Test 2	0.65	0.64
Test 3	0.70	0.65

We also ran correlation analyses to check the test/re-test reliability. Tables 4 & 5 show the correlations of all paper and computer tests with each other and with the reading section of the TOEFL as a concurrent validity indicator:

Tables 4 & 5: Correlations of all paper and computer tests with each other

		paper-based test 2	**paper-based test 3**
paper-based test 1	Pearson Correlation	.243**	.187*
	Sig. (2-tailed)	.002	.016
paper-based test 2	Pearson Correlation		.391**
	Sig. (2-tailed)		.000

		computer-based test 2	**computer-based test 3**
computer-based test 1	Pearson Correlation	.464**	.423**
	Sig. (2-tailed)	.000	.000
computer-based test 2	Pearson Correlation		.466**
	Sig. (2-tailed)		.000

Table 6: Correlations of all paper and computer tests with TOEFL

		Reading section of TOEFL
paper-based test 1	Pearson Correlation	.318**
	Sig. (2-tailed)	.000
paper-based test 2	Pearson Correlation	.354**
	Sig. (2-tailed)	.000
paper-based test 3	Pearson Correlation	.370**
	Sig. (2-tailed)	.000
computer-based test 1	Pearson Correlation	.355**
	Sig. (2-tailed)	.000
computer-based test 2	Pearson Correlation	.386**
	Sig. (2-tailed)	.000
computer-based test 3	Pearson Correlation	.433**
	Sig. (2-tailed)	.000

The results indicate that all paper tests significantly correlated with each other and with the reading section of TOEFL. Furthermore, there is a significant correlation between all the computer-based tests with the TOEFL reading section. The overall correlations favoured the computer-based tests. A further validity check was to correlate the computer-based tests scores with the constructs and mediators under investigation i.e. computer familiarity, computer attitude, and testing mode preference. Table 7 demonstrates the correlation results:

Table 7: Correlations of all computer-based tests with all study variables

		Computer familiarity scale	**Computer familiarity tasks**	**Computer liking Scale**	**Testing mode pref.**
computer-based test 1	Pearson Correlation	.133	-.129	.105	.050
	Sig. (2-tailed)	.088	.098	.177	.522
computer-based test 2	Pearson Correlation	.048	-.076	.056	.010
	Sig. (2-tailed)	.541	.330	.475	.901
computer-based test 3	Pearson Correlation	.058	-.135	.140	.077
	Sig. (2-tailed)	.460	.081	.071	.322

From Table 7, we find that there is no significant correlation between the computer-based tests, which is our interest here, and the other moderators such as computer attitude and testing modes preference, and other constructs such as computer familiarity with its two measures i.e. the scale and the computer tasks. This indicates that these variables have no effect on the scores of computer-based tests and consequently there is no impact on the overall validity of these tests. To sum up, there was no significant effect of the testing mode on the overall reliability and validity of the tests.

This consolidated conclusion answers our first research question. It is essential to mention that our findings for this research question agree with other relevant studies in the field. Our findings about the effect of testing mode on test reliability and validity match some of the results in related research. For example, the findings of Olsen et al. (1989) confirmed that paper-based and computer-based as well as the computer adaptive tests of the mathematical items of California Assessment Program (CAP) yielded equivalent scores. Boo (1997) also found that testing modes did not influence the reliability of tests, and other constructs such as computer familiarity did not appear to be part of the construct measured by the computerized tests. This means that neither the testing mode nor the assumed construct had an impact on the overall reliability or construct validity of the tests. Furthermore, the findings of Choi et al., (2003) supported the comparability of PBTs and CBTs of the Test of English Proficiency prepared by Seoul National University (TEPS).

In order to establish PBT and CBT comparability, it is essential to ensure that scores produced by both forms are a true measure of the same construct. Thus, in our study, we examined the relationship between the construct of computer familiarity and the subjects' performance on CBT. Since we used two measures for the construct of computer familiarity, both measures will be used in our analysis to get more reliable results upon which we can draw valid conclusions. To answer this question, we examined the relationship between computer familiarity and participants' performance on computer-based tests. First, we ran correlations between the two computer familiarity indicators and the computer-based tests. Table 8 shows the correlation results:

Table 8: Correlation of all computer-based tests with computer familiarity measures

		Computer Familiarity Scale	**Computer tasks**
Mean of computer-based tests	Pearson Correlation	.096	-.138
	Sig. (2-tailed)	.215	.075

This indicates that there was no significant correlation between the subjects' performance on computer-based tests and their computer familiarity scale. Furthermore, the correlation between the second measure of computer familiarity i.e. computer tasks and the computer-based tasks, was not significant. These results indicate that there is no significant relationship between computer familiarity and performance on computer-based tests. We also performed repeated measure ANOVA to explore the relationship between computer familiarity and participants' performance on computer-based tests. The results are presented in Table 9:

*Table 9: ANOVA results of interaction of tests*modes*computer familiarity measures*

Source		**df**	**F**	**Sig.**
tests * modes * computer familiarity scale	Sphericity Assumed	2	1.691	.186
tests * modes * computer familiarity tasks	Sphericity Assumed	2	2.710	.068

Computer familiarity construct with its two measures i.e. scale and tasks do not have a significant effect on computer-based tests. All these results answer our second research question. Our findings here are in line with other research findings in the literature (Powers & O'Neill, 1992; Vispoel, et al., 1994; Boo, 1997; Taylor, et al., 1999; Fulcher, 1999; Higgins, et al., 2005).

When examining the comparability of PBT and CBT, the participants' computer attitude should be taken into account. Hence, the third research question deals with this issue. To answer this question, we used correlation and repeated measure ANOVA. Table 10 summarises the results:

Table 10: Correlation of computer attitude measure and computer-based tests

		Mean of Computer-based Tests
Computer Attitude Scale	Pearson Correlation	.121
	Sig. (2-tailed)	.118
	N	167

There is no significant correlation between computer attitude and the participants' performance on computer-based tests. We used repeated measure ANOVA to examine the effect of computer attitude on performance on computer-based tests. Results are presented in Table 11:

Table 11: Repeated measure ANOVA

Source		**df**	**F**	**Sig.**
tests * modes * computer attitude	Sphericity Assumed	2	.122	.885

These results indicate that computer attitude has no significant interaction with the tests and the modes. To sum up, we found no significant correlation between computer attitude and subjects' performance. Moreover, the repeated measure ANOVA confirmed that participants' computer attitude has no significant effect on their performance. All these results answer our third research question. These results go with other research findings such as those of Powers & O'Neill (1992).

Examining the relationship between testing mode preference and performance when conducting a PBT and CBT comparability study is essential. To answer our four question, the subjects' responses to a simple question in Questionnaire One, i.e. *Would you prefer taking test on: paper – no difference - computer*, were correlated with their mean scores on computer-based tests. Our coding for respondents' answers was 1= on paper, 2= no difference, 3= on computer. Table 12 shows the results of these correlations

Table 12: Correlations of Pre-CBT testing mode preference and computer-based tests

		Mean of Computer-based Tests
Pre-CBT testing mode preference	Pearson Correlation	.054
	Sig. (2-tailed)	.490
	N	167

Apparently, there is no significant correlation between participants' pre-CBT testing mode preference and their performance on either testing mode. We also performed multiple comparisons between preference groups using one-way ANOVA to examine the relationship between the prior testing mode preference and performance on computer-based tests. Table 13 shows the one-way ANOVA results:

Tables 13: Multiple comparisons of Pre-CBT preference groups

(I) Would you prefer taking tests	(J) Would you prefer taking tests	Sig.
On paper	No difference	.175
	On computer	.925
No difference	On paper	.175
	On computer	.498
On computer	On paper	.925
	No difference	.498

Dependent Variable: Mean of computer-based tests

These results also support the absence of the interaction between pre-CBT testing mode preference and performance on computer-based tests. Previous results confirmed that there is no significant relationship between the participants' testing mode preference and their performance on computer-based tests.

To seek an answer to the fifth question concerning the influence of exposure to CBT on testing mode preference, we used descriptive statistics. The following table shows the descriptive statistics for each group:

Table 14: Descriptive statistics of groups according to their Pre-CBT preference

Would you prefer		N	Mean	Standard deviation
On paper	Paper based test 1	68	76.08	13.689
	Paper based test 2	68	64.08	20.164
	Paper based test 3	68	64.12	18.322
	Computer based test 1	68	72.65	16.172
	Computer based test 2	68	59.98	19.866
	Computer based test 3	68	62.84	14.854
	Valid N (listwise)	68		
No difference	Paper based test 1	47	81.42	10.397
	Paper based test 2	47	67.63	15.539
	Paper based test 3	47	70.78	18.575
	Computer based test 1	47	78.16	14.560
	Computer based test 2	47	66.87	15.792
	Computer based test 3	47	64.11	14.682
	Valid N (listwise)	47		
On computer	Paper based test 1	52	75.64	13.260
	Paper based test 2	52	68.54	16.511
	Paper based test 3	52	70.26	17.686
	Computer based test 1	52	74.10	14.799
	Computer based test 2	52	59.89	20.062
	Computer based test 3	52	65.64	16.404
	Valid N (listwise)	52		

From the descriptive statistics in Table 14, it is quite evident that the participants who preferred paper tests and the second group which did not mind taking the test on either mode, did better on the paper tests. The last group, which selected CBT as its preferred testing mode, did better on paper tests, however. The overall results answer quite negatively the fifth research question. These findings showed that all participants did better on paper-based tests, though their performance was not particularly high.

The overall results confirm that there is neither significant effect nor interaction between prior testing mode preference and performance on either of the testing modes. This is an additional indication that testing mode preference does not affect test validity. Nevertheless, subjects tend to perform better on paper-based tests than on computer-based tests regardless of their testing mode preference. We would argue that this is due to the novelty of CBT in the target context and might be attributed to some of the reasons brought up by subjects in the subsequent interviews, such as the eye fatigue, page scrolling, and text display. This should be taken into consideration in the target context if CBT is to be implemented. A possible solution for this is to adjust the pass/fail cut-off points. The overall findings here agree with Fulcher (1999) who examined the relationship between students' attitudes about computer-based testing and performance on web-based testing and found no significant impact of participants' attitude on their CBT scores. It is the same conclusion that Russell (1999) arrived at when investigating the examinees' preference for writing on paper or using a keyboard when they were doing science, mathematics and language arts tests.

In terms of our sixth research question we aimed to examine the testing mode preference before and after participants were exposed to CBT to investigate the impact of exposure to CBT on subjects' testing mode preference. Therefore, we divided this part of our research into two phases: Pre-CBT testing mode preference and Post-CBT testing mode preference. To measure each of these, we asked the participants about their testing mode preference before and after exposure to CBT in the First and Second Questionnaires. Table 15 shows the frequencies of participants' responses BEFORE exposure to CBT.

Table 15: Frequency table of responses to the Pre-CBT testing mode preference

Valid:	Frequency	Percent	Valid Percent	Cum. Percent
On paper	68	20.4	40.7	40.7
No difference	47	14.1	28.1	68.9
On computer	52	15.6	31.1	100.0
Total	167	50.0	100.0	

From Table 15, we can see that 40.1% preferred to take the test on paper, 28.1% did not mind taking the test in either mode while 31.8% opted for computers as their preferred mode of testing. Participants justified their preferences differently. For instance, for those who opted for computers as their preferred testing mode, their motives ranged from CBT being an innovation in the assessment system, the accuracy of CBT, and time saving, to the enjoyment of CBT and its ease of recording and changing answers.

On the other hand, participants who chose paper as their preferred testing mode attributed that to the following factors: lack of keyboarding skills, ease and comfort, intolerable CBT technical faults, and familiarity with this type of assessment. They also preferred paper-based testing because reading the questions and recording the answers is easier, as is highlighting the text, and it does not cause any eye fatigue. For neutral participants, they shared the same view of each testing mode advocate and thus they have no reservations about either mode. After the subjects had finished both paper- and computer-based tests, they were asked once more about which test they would prefer to take again. Table 16 presents the distribution of participants in the three categories:

Table 16: Frequency table of responses to Post-CBT testing mode preference

Valid:	Frequency	Percent	Valid Percent	Cum. Percent
On paper	58	34.7	34.7	34.7
No difference	22	13.2	13.2	47.9
On computer	87	52.1	52.1	100.0
Total	167	100.0	100.0	

These tables show that only 34.7% still preferred paper-based tests while only 13.2% prefer taking their tests in either mode. The greater percentage (52.1%) was those who chose CBT as their preferred mode of testing. When rationalizing their preferences, participants gave almost the same reasons they had already given to the previous preference question in Questionnaire One. The results indicate that there is a significant distribution of participants in categories and this is particularly apparent in the second and third categories. We can conclude from the previous results that the number of subjects who preferred PBT and those who preferred taking tests in either mode have changed to favour those who chose CBT as their preferred testing mode. However, identifying the attributes of this significant alteration of preference was crucial. Therefore, we conducted retrospective interviews with a random sample of those who had changed their testing mode preference from PBT to CBT, from CBT to PBT, and those who did not mind either testing mode. It was essential to know if the participants had changed their preference solely because of the experience itself.

Responses were collected from 23 participants. 53% of the participants changed their testing mode preference from either PBT or neutral to CBT. Some subjects justified their prior paper test preference by having no prior CBT experience. Some also attributed this to past unpleasant CBT experiences such as some boring computer courses. Furthermore, being accustomed to paper tests, the novelty of CBT was another reason for some subjects. Eye fatigue was a major concern for one participant from his prior experience with the daily use of computers. However, after these participants had been involved in the CBT experience, they entirely shifted from PBT as their preferred testing mode to CBT. They found CBT more comfortable, more enjoyable, and time saving. Ease of changing answers, reading the passage and questions, as well as being able to navigate through the text and the questions were very attractive features of CBT that influenced the subjects' testing mode preference. Participants also liked the display of the passages and the questions which was an innovation for them in test taking experience as well as a shift from the classical testing mode i.e. PBT to the new technological one, CBT.

On the other hand, it is also important to examine the other group who altered their preference from either CBT or neutral to PBT. This group of participants attributed this change to unfamiliarity with computers and technology. Although they felt that CBT is more comfortable, enjoyable and saves time, certain issues led these participants to change their preference to subsequently favour PBT. Their evaluation of CBT was negative since it was their first unpleasant and uncomfortable CBT experience. Physical and psychological problems caused by CBT, such as eye fatigue and boredom, were other motives behind their preference alteration. Some technological issues such as text display, scrolling, and a test indicator of time remaining affected the students' preference for CBT. Subjects stressed these concerns and asserted that they would change their preference to CBT if their concerns with CBT are to be taken into account and overcome in future CBT experiences.

We also investigated the impression or feelings that all participants had developed about CBT features after being involved in our study. Table 17 summarizes the results:

Table 17: Frequencies of responses of posterior preference to features of PBT and CBT

Questions	N= 167		
	Options		
	On paper	No Difference	On computer
In which test was reading passages easier to navigate through?	63	19	85
In which test was reading passages easier to read?	88	23	56
In which test was the text in the items easier to read?	52	44	71
Which test was less fatiguing?	47	28	92

Questions	N= 167		
	Options		
	On paper	No Difference	On computer
In which test was it easier to record answers?	39	32	96
In which test was it easier to change answers?	9	7	151
Which test were you more likely to guess the answer in?	29	88	50
Which test was more comfortable to take?	63	20	84
In which test would you be more likely to receive the same score if you took it a second time?	52	70	45
Which test was more enjoyable to take?	20	20	127
Which test more accurately measured your reading comprehension skills?	85	44	38

Table 17 indicates that more than half of the participants developed a positive attitude towards the majority of CBT features. For instance, it was easier for 51% of the subjects to navigate through the passages on computer than on paper and 43% found it easier to read the test items on the computer than on paper. Moreover, about 55% felt it less fatiguing to take a test on computer than on paper. 57% and 90% respectively found recording and changing the answers easier on computer. Not only that but also 50% of the subjects felt more comfortable when taking the test on computer than on paper and 76% enjoyed it. Yet, about 52% found it easier to read the text on paper than on screen. One interesting finding is the percentages of the last question about the accuracy of the two modes for measuring the reading comprehension skill of the participants. About 50% of the subjects think that PBT measures their comprehension skills accurately while only 21% think that CBT is better in this respect. I would argue that these percentages do not contradict their post-CBT preferences where 34% chose PBT again while 52% went for CBT as their preferred testing mode.

Conclusion

This study aimed to measure the comparability of both paper- and computer-based L2 reading achievement tests. It also investigated the relationship between several factors i.e. computer familiarity, computer attitude, and testing mode preference and performance on computer-based tests. Thus far, we have found that the testing mode has no significant effect on the overall validity and reliability of the tests. We also reached a point where we can assert that the construct of computer familiarity has no influence on students' performance on computer-based tests. In addition, the other moderators such as computer attitude and testing mode preference do not have any critical impact on the overall students' performance on computer-based tests. Since our study still has a qualitative part in its early stages of coding and analysis, no clear picture can yet be offered about the final findings and results of this study. Therefore, final solid conclusions cannot yet be drawn out of the available results due to the incomplete data analysis. This study is still ongoing and it is hoped that by Spring 2009, the complete results and findings will be ready for publication.

References

A. Bugbee. 1996. The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education*, vol. 28 (3) pp. 282-300.

A. Durndell & P.Lightbody. 1993. Gender and computing: Change over time? *Computers in Education*, 21, pp. 331-336.

B. Green, R. Bock, L. Humphreys, R. Linn, & M. Reckase. 1984. Technical guidelines for assessing compurerised adaptive tests. *Journal of Educational Measurement*, 21, pp. 374-359.

B. Loyd & C. Gressard. 1984. The effect of sex, age, and computer experience on computer attitudes. *AEDS Journa*l, 40, pp. 67-77.

B. Whitworth. 2001. *Equivalency of paper-and-pencil tests and computer-administered tests*. Unpublished dissertation, University of North Texas.

C. J. Weir. 2005. *Language Testing and Validation: An Evidence-Based Approach*. Palgrave Macmillan: NY, USA.

C. Parshall, J. Spray, J. Kalohn, & T. Davey. 2002. *Practical Considerations in Computer-Based Testing*. New York: Springer.

C. Taylor, I. Kirsch, D. Eignor, J. Jamieson. 1999. Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning,* vol. 49 (2). pp. 219-274.

C. Taylor, J. Jamieson, D. Eignor, I. Kirsch. 1998. The relationship between computer familiarity and performance on computer-based TOEFL test tasks. *TOEFL Research Reports. Report 61.* Princeton, NJ, USA: Educational Testing Services.

Carol Chapelle & Dan Douglas. 2006. *Assessing language through computer technology*. Cambridge. UK: CUP.

Carol Chapelle. 2001. *Computer Applications in Second Language Acquisition: Foundations for teaching, Testing and Research*. CUP.

D. Dalton. 1994. What do others know about CD-ROMs, LANs, modems, and more. A survey for staff training. *The Book Report*, 12, 19.

D. Eignor, C. Taylor, I. Kirsch, and J. Jamieson. 1998. Development of a scale for assessing the level of computer familiarity of TOEFL examinees. *TOEFL Research Reports, Report 60*. Princeton, NJ, USA: Educational Testing Services.

D. Powers & K. O'Neill. 1993. Inexperienced and anxious computers users: coping with a computer-administered test of academic skills. *Educational Assessment*, 1, pp. 153-173.

D. Stephen & F. Rowland. 1993. Initial IT training in Departments of Information and Library Studies in the British Isles: A survey of student views. *Education of Information*, 11, pp. 189-204.

F. Millar & N. Varma. 1994. The effect of psychological factors on Indian children's attitudes toward computers. *Journal of Educational Computing Research*, 10, pp. 223-238.

F. Vincino & K. Moreno. 1988. Test-taker's attitudes toward and acceptance of a computerised adaptive test. *Paper presented at the annual meeting of the American Educational Research Association*, New Orleans, USA.

G. Fulcher. 1999. Computerizing an English language placement test. *ELT Journal*, 53(4), pp. 289-299.

G. Marcoulides. 1988. The relationship between computer anxiety and computer achievement. *Journal of Educational Computing Research*, 4, pp. 151-158.

I. Choi, K. Kim, and J. Boo. 2003. Comparability of a paper-based language test and a computer-based language test. *Language Testing*, vol. 20 (3) pp. 295-320.

I. Kirsch, J. Jamieson, C. Taylor, and D. Eignor. 1998. Computer familiarity among TOEFL examinees. *TOEFL Research Reports, .Report 59*. Princeton, NJ, USA: Educational Testing Services.

J. Boo .1997. *Computerized versus paper-and-pencil assessment of educational development: Score comparability and examinee preferences*. Unpublished dissertation, University of Iowa.

J. C. Alderson. 1991. Dis-sporting life. Response to Alistair Pollit's paper. In Alderson and North (eds.). *Language testing in the 1990s.* pp. 60-67. London: Macmillan.

J. C. Alderson. 2000. Technology in Testing: the Present and the Future. *System*, 28(4), pp. 593-603.

J. Geissler & P. Horridge. 1993. University students' computer knowledge and commitment to learning. *Journal of Research on Computing in Education*, 25, pp. 347-365.

J. Lee. 1986. The effect of mode of past computer experience on computerized aptitude performance. *Educational and Psychological Measurement*, 46, pp. 727-733.

J. Mazzeo & L. A. Harvey. 1988. The equivelance of scores from automated and conventional education and psychological tests: A review of the literature. *(Report No. CBR 87-8, ETS RR 88-21).* Princton, NJ: Educational Testing Services.

J. Mazzeo, B. Druesne, P. Raffeld, K. Checketts & A. Muhlstein. 1991. Comparability of computer and paper-and-pencil scores for two CLEP general examinations. *(College Board Report 91-5).* Princton, NJ: ETS.

J. Olsen, D. Maynes, D. Slawson & K. Ho. 1989. Comparison of paper-administered, computer-administered and computerized achievement test. *Journal of Educational Computing Research*, Vol.5, pp. 311-326.

M. Chalhoub-Deville & C. Deville. 1999. Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics*, 19, pp. 273-99.

M. Hicks. 1989. The TOEFL computerized placement test: adaptive conventional measurement. *(ETS Reports No. 89-12)*. Princeton, NJ: Educational Testing Services.

M. J. Bresolin. 1984. *A comparative study of computer administration of the Minnesota Multiphasic personality Inventory in an inpatient psychiatric setting*. Unpulished doctoral dissertation, Loyola University, Chicago, USA.

M. Pomplun, S. Frey, & D. BECKER. 2002. The score equivalence of paper –and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement*, Vol. 62 No. 2, pp. 337-354.

M. Russell, & W. Haney. 1997. *Testing writing on computers: An experiment comparing students performance on tests conducted via computer and via paper-and-pencil.* Educational Policy Analysis Archive, vol. 5 (3).

M. Russell, A. Goldberg & K. O'Conner. 2003. Computer-based testing and validity: a look back into the future. *Assessment in Education*, Vol. 10-3, pp. 279-293.

M. Russell. 1999. *Testing on computers: A follow-up study comparing performance on computer and on paper.* Educational Policy Analysis Archive, vol. 7 (20).

O. Chrismas. 1992. Use of technology by special education personnel. Lansing, MI: Michigan Department of Education, Bureau of information Management.*(ERIC Document Reproduction Service No. ED 350 743).*

O. Jegede & P. Okebukola. 1992. Adopting technology in third world classrooms: students' viewpoint about computers in science teaching and learning. *Journal of Educational Technology Systems*, 20, pp. 327-335.

P. Hofer & B. F. Green. 1985. The challenge of competence and creativity in computerized psychological testing. *Journal of Counseling and Clinical Psychology*, 53, pp. 826-838.

P. Paek. 2005. Recent trends in comparability studies. *Pearson Educational Measurement Research Reports. Research Report 05-05*. Pearson Educational Measurement. USA.

R. Okinaka. 1992. Sex differences in computer backgrounds and attitudes: a study of teachers and teacher candidates. San Bernadino, CA: California State University, instructional Technology Program. *(ERIC Document Reproduction Services No. ED 353952).*

T. Harrel, M. Honaker, M. Hetu & J. Oberwager.

1987. Computerized versus traditional administration of the multidimensional aptitude battery-verbal scale: an examination of reliability and validity. *Computers in Human Behavior*, 3, pp. 129-137.

T. Levin & C. Gordon. 1989. Effect of gender and computer experience on attitudes towards computers. *Journal of Educational Research*, 5, pp. 69-88.

T. Ward, S. Hooper & K. Hannafin. 1989. The effects of computerized tests on th performance and attitudes of college students. *Journal of Educational Computing Research*, 5, pp. 327-333.

V. Bunderson, D. Inouye & J. Olsen. 1989. The four generations of computerized educational measurement. In R. L. Linn (Ed). *Educational Measurement*. pp. 367-407. Phoenix, AZ: Oryx Press.

Y. Sawaki. 2001. *Comparability of conventional and computerized tests of reading in a second language. Language Learning & Technology*, vol.5 (2) pp. 38-59.