

## **Computer-based testing vs. paper-based testing: a comprehensive approach to examining the comparability of testing modes**

**Saad Al-Amri<sup>1</sup>**

### **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. 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 EFL tests in medical colleges in Saudi Arabia as they are the main audience in the target context. This paper reports on the results of the quantitative instruments of the study i.e. the tests and the questionnaires and the relevant interviews. Although we found a significant difference between the mean scores on both modes, none of the factors examined had an influence on students' performance when doing the computer-based tests. The researcher is still in the process of analyzing the qualitative data and thus this paper focuses mainly on the quantitative part of the research.

### **1 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 aspects of the assessment, which include the concepts of validity and reliability.

---

<sup>1</sup> salamr@essex.ac.uk

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, 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 in 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 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 other great advantages of CBT. Collecting different performance data such as latency<sup>2</sup> 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, which led many scholars to suggest conducting systematic studies to check equivalency and comparability of paper-based tests and computer-based tests (Parshall et al., 2002). For example, students need some degree of computer literacy in order to avoid the mode effect on computer-based testing (Alderson, 2000).

---

<sup>2</sup> An anonymous reviewer requested clarification of the term 'latency'. However, this term was not defined in the main source, but the author assumes it refers to the response time latency for test items.

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”.

One of the main contributing factors that should be examined when conducting 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 of vocabulary, ability to interpret literary materials, and ability to do quantitative thinking. Moreover, Taylor et al. (1999) found no evidence of an undesirable effect of computer familiarity on students’ performance on computer-based TOEFL test. 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).

Essential additional test taker characteristic that might affect his/her performance with regard to CBT is computer attitude. 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 & Lancaster, 1987, cited in Levin & 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 & Moreno, 1988; Levin & Gordon, 1989; Bruke et al., 1987; Powers & O'Neill, 1992 and Boo, 1997).

Therefore, this study aims to measure the comparability of computer-based and paper-based tests of institutional multiple choice reading achievement 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, (though test taking strategies are not discussed in this article), 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.

## **2 The study**

### **2.1 Research questions**

This study 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?
5. Will participants' posterior testing mode preference be influenced by exposure to CBT? If so, why?
6. After exposure to CBT, will subjects develop a more positive attitude towards PBT or CBT features?

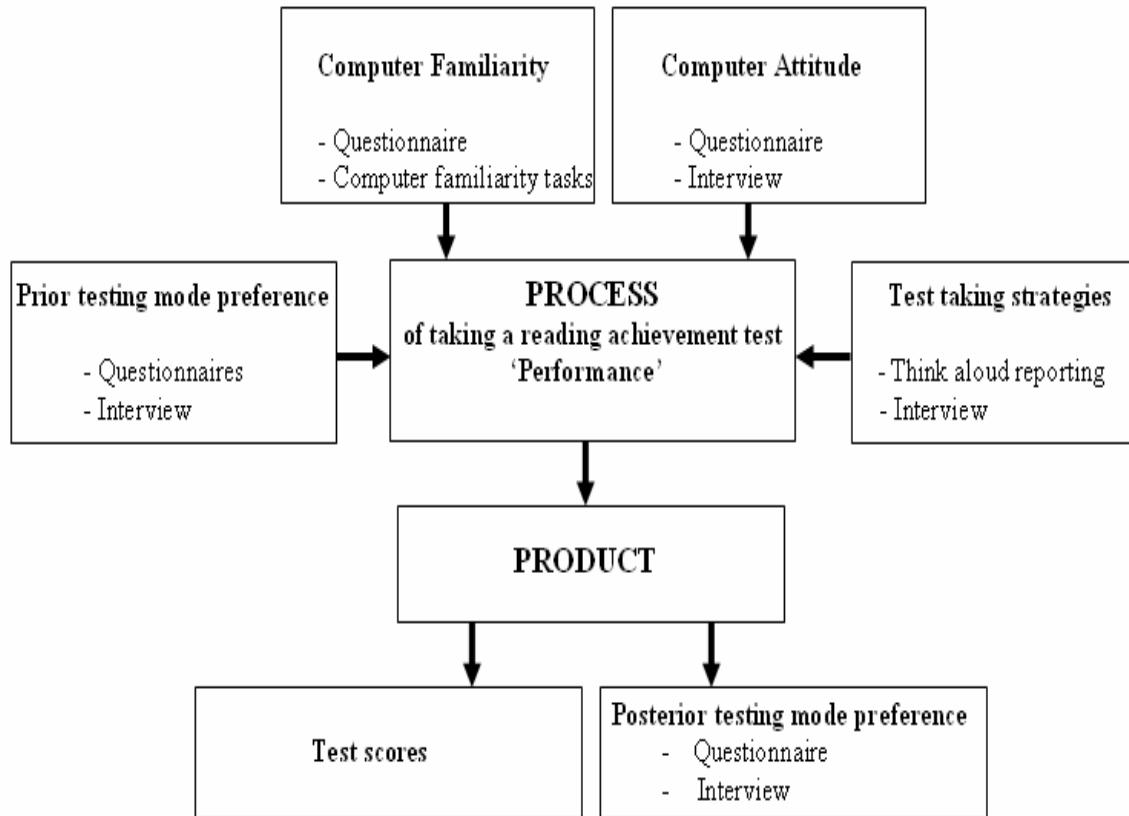
### **2.2 Methodology**

This study used triangulation of a variety of data sources from students doing both paper and computer-based reading tests in order 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. Figure 1 shows a model of supposed interaction steps between the key variables and how each is measured<sup>3</sup>:

---

<sup>3</sup> Test taking strategies are not discussed in this paper.

**Figure 1**



### 2.3 Subjects

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

### 2.4 Study instruments

The study employed the following instruments:

- The reading section of the TOEFL test to measure reading proficiency (with permission from ETS).

- Two computer familiarity exercises designed by the researcher to measure, in a direct way, the participants' computer familiarity.
- Two questionnaires measuring computer familiarity (21 items), computer attitude (8 items), prior and posterior testing mode preference (2 items) and attitudes to CBT features (17 items) designed by the researcher with the help of questionnaires available in the literature.
- Three institutional reading multiple choice achievement tests (44 MCQs in total) made available to the researcher by the target institution.
- Semi-structured interviews conducted after administering the second questionnaire and the verbal reports.
- Think aloud verbal report protocols to gain insights into mental processes used by subjects while taking the tests on both testing modes (not covered in this article).

## 2.5 Procedure

All instruments were administered over three months by the researcher in class time with appropriate consent from both the university and the subjects. 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 and computer attitudes 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 QuestionMark system *Perception*. Prior to testing sessions, there was a tutorial for all subjects to familiarize them with the new testing mode i.e. CBT and reduce the possible testing mode familiarity effect. 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. All subjects took the three tests i.e. (test 1, 2, and 3) where each subject took the same test once on paper and once on computer. 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.

## 2.6 Data analysis

Not all variables and their relevant instruments can be presented and discussed in this article. This article will focus only on the results of the tests on both modes and their relationship with the explanatory variables i.e. computer familiarity, computer attitudes, and prior and posterior testing mode preference. The process variable i.e. (test taking strategies) and its instruments will not be presented and discussed as the coding and analyses of interviews and think aloud have not yet been completed.

## 3 Results and discussion

### 3.1 Validity and reliability (RQ 1)

First, starting with construct validity, the test score 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. Deviation
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 paired-sample 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

Tests		t	p
Pair 1	paper-based test 1 - computer-based test 1	2.42	.016
Pair 2	paper-based test 2 - computer-based test 2	3.28	.001
Pair 3	paper-based test 3 - computer-based test 3	2.99	.003

We would argue that these significant differences are attributed to the low number of test items on each test (15 items in test 1, 14 items in test 2, and 15 items in test 3). Moreover, any small difference in a large sample size (167 subjects) may be significant. It is quite obvious from table 1 that the difference between the means of each test is not vast so validity is not impaired. Nevertheless, there is a crucial issue which lies in the issue of pass/fail marks. In the target context, the subjects are required to obtain 60% in the test in order to pass. Our results revealed that in test 1, for instance, (10%) failed on the computer test but passed the same test on paper. The percentage increased in test 2 where (24%) failed the computer-based version and passed its paper-based counterpart. Conversely, (16%) were not able to pass test 3 on paper; however, they passed its computerized version. Yet, there was no subject who passed all the tests on paper and failed all the tests on computer which could be clearly attributed to the testing mode. Furthermore, most of those subjects who failed some of the computer-based tests were able to pass at least one computer-based test and failed the same test on paper. In general, subjects tended to perform better on paper-based tests than on computer-based tests. This might be 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 eye fatigue, difficulty of page scrolling, and unfamiliarity of 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.

To examine the effect of the testing mode on reliability, we examined the internal consistency (Cronbach's alpha) of each test on each mode. Table 3 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

Although these reliability coefficients are not as high as one would like (maybe because the tests are not professionally made), but they are very similar between the modes. We also ran correlation analyses to check the test/re-test reliability. Table 4 shows the correlations of each paper-based test with its computer-based version and table 5 with the reading section of the TOEFL as a concurrent validity indicator:

Table 4: Correlations of all paper and computer tests with each other

		computer-based test 1	computer-based test 2	computer-based test 3
paper-based test 1	r	.456**		
	Sig.	.000		
paper-based test 2	r		.525**	
	Sig.		.000	
paper-based test 3	r			.529**
	Sig.			.000

Table 5: 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 each paper-based test significantly correlated with its computerised version. Furthermore, there is a significant correlation between all the paper-based and computer-based tests with the TOEFL reading section. The overall correlations favoured the computer-based tests. However, these correlations are quite low due to the tests that are institutional compared with the TOEFL which is a professionally made test.

A further validity check was to correlate the computer-based tests scores with the variables potentially affecting test scores (see section 1) under investigation i.e. computer familiarity, computer attitude, and testing mode preference. Table 6 shows the correlation results:

Table 6: Correlations of all computer-based tests with the study variables

		Computer Familiarity Scale	Computer familiarity tasks	Computer attitude Scale	Prior testing mode preference
computer-based test 1	r	.133	-.129	.105	.050
	Sig.	.088	.098	.177	.522
computer-based test 1	r	.048	-.076	.056	.010
	Sig.	.541	.330	.475	.901
computer-based test 1	r	.058	-.135	.140	.077
	Sig.	.460	.081	.071	.322

From Table 6, we find that there is no significant correlation between the computer-based test scores, which is our interest here, and the construct irrelevant variables such as computer familiarity with its two measures i.e. the scale and the computer tasks, computer attitude and testing modes preference. This indicates that these variables have no effect on the scores of computer-based tests and consequently there is no impact on the construct validity of these CBT tests. Our findings here are in line with the findings of Boo (1997).

To sum up, although there was a significant difference between the scores on the two modes, this difference was not a result of the testing mode effect (see p.8). Moreover, other analyses confirmed that 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. 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 construct-irrelevant variable 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 computer familiarity and computer attitude had an impact on the overall reliability or construct validity of the tests. Furthermore, the findings of Choi et al. (2003) supported the scores comparability of PBTs and CBTs of the Test of English Proficiency prepared by Seoul National University (TEPS).

### **3.2 Computer familiarity and CBT (RQ 2)**

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 which is the second research question in this article. Thus, in our study, we examined the relationship between the construct-irrelevant variable i.e. computer familiarity and the subjects' performance on CBT. Since we used two measures 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 RQ2, we examined the relationship between computer familiarity and participants' performance on computer-based tests. First, we ran correlations between the two computer familiarity measures and the mean of computer-based tests. There was no significant correlation between the subjects' performance on computer-based tests and their computer familiarity scores ( $r=.09$ ,  $p=.21$ ). Furthermore, the correlation between the second measure of computer familiarity i.e. computer tasks and the mean of computer-based tests, was non significant ( $r=-.13$ ,  $p=.07$ ). 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 7:

Table 7: ANOVA results of interactive effect of modes\*computer familiarity measures on test scores

Source		df	F	Sig.
modes * computer familiarity scale	Sphericity Assumed	1	.208	.649
modes * computer familiarity tasks	Sphericity Assumed	1	1.769	.185

The computer familiarity variable with its two measures i.e. questionnaire scale and practical 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).

### 3.3 Computer attitudes and CBT (RQ 3)

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<sup>4</sup> and repeated measure ANOVA. The results showed that there is no significant correlation between computer attitudes 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 8:

Table 8: Repeated measure ANOVA

Source		df	F	Sig.
modes*computer attitude	Sphericity Assumed	1	1.060	.305

These results indicate that computer attitude has no significant interaction effect with 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 on CBT.

---

<sup>4</sup> This correlation is with the mean of the CBTs while the correlations on p.11 are with separate tests.

All these results answer our third research question. These results fit with other research findings such as those of Powers & O’Neill (1992). RQ4 is discussed next.

### 3.4 Testing mode preference and performance on PBT and CBT (RQ 4)

Examining the relationship between testing mode preference and performance when conducting a PBT and CBT comparability study is essential (McDonald, 2002). To answer our fourth question, the subjects’ responses to a simple question in Questionnaire One, i.e. *Would you prefer taking test on: paper – no difference - computer*, was 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 9 shows the results of these correlations.

Table 9: Correlations of Pre-CBT testing mode preference and computer-based tests

		Mean of Computer-based Tests	Mean of paper-based Tests
Prior testing mode preference	Pearson Correlation	.054	.130
	Sig. (2-tailed)	.490	.095
	N	167	167

Apparently, there is no significant correlation between participants’ prior testing mode preference and their performance on either testing mode. We also performed repeated measure ANOVA to examine the relationship between the prior testing mode preference and performance on computer-based tests. Table 10 shows the ANOVA results:

Tables 10: Repeated measure ANOVA

Source		df	F	Sig.
modes * prior testing mode preference	Sphericity Assumed	1	.448	.504

These results also support the absence of the relationship between prior testing mode preference and performance on paper and computer-based tests and confirm other studies.

### 3.5 Exposure to CBT and testing mode preference (RQ 5)

Our fifth research question 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. To measure that, we asked the participants about their testing mode preference before and after exposure to CBT in the First and Second Questionnaires. Table 11 shows the frequencies of participants' responses BEFORE and AFTER exposure to CBT:

Table 11: Prior testing mode preference \* posterior testing mode preference crosstabulation

			Prior testing mode preference			Total
			On paper	No difference	On computer	
Posterior testing mode preference	On Paper	Count %	32 19.2%	15 9.0%	11 6.6%	58 34.7%
	No Difference	Count %	8 4.8%	11 6.6%	3 1.8%	22 13.2%
	On Computer	Count %	28 16.8%	21 12.6%	38 22.8%	87 52.1%
Total		Count %	68 40.7%	47 28.1%	52 31.1%	167 100.0%

From Table 11, we can see that 40.7% preferred to take the test on paper, 28.1% did not mind taking the test in either mode while 31.1% opted for computers as their preferred mode of testing. Participants justified, in the first questionnaire, 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. Neutral participants shared the same view of each testing mode and thus had no reservations about either mode. After the subjects had finished both paper- and computer-based tests, they were asked once more about on which testing mode they

would prefer to take the test again. According to table 13, 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, in the second questionnaire, participants gave almost the same reasons they had already given to the previous preference question in Questionnaire One. We used Chi-square test to examine the significance of difference between the prior and posterior preference groups. Table 12 shows the Chi-square results:

Table 12: Chi-square results

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	18.290(a)	4	.001
Likelihood Ratio	17.977	4	.001
Linear-by-Linear Association	11.423	1	.001
N of Valid Cases	167		

The results indicate that there is a significant change between subjects' prior and posterior testing mode preference and this is particularly apparent in the second and third categories. We can conclude from table 13 and 14 that the number of subjects who preferred PBT and those who preferred taking tests in either mode have changed significantly 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 and from CBT to PBT or to neutral. It was important 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 experiences such as 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.

It is quite evident that there was a significant change between the subjects preference before and after they were exposed to CBT. The overall results confirm that there is no significant relationship 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. The overall findings here agree with Fulcher (1999) who examined the relationship between students' attitudes to 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.

### 3.6 Attitudes developed towards PBT and CBT features (RQ 6)

The sixth research question investigated the attitudes that all participants had developed about PBT and CBT features after being involved in our study. Table 13 summarizes the results:

Table 13: Frequencies of responses reflecting posterior attitudes 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?	37.7	11.4	50.9
In which test was reading passages easier to read?	52.7	13.8	33.5
In which test was the text in the items easier to read?	31.1	26.3	42.5
Which test was less fatiguing?	28.1	16.8	55.1
In which test was it easier to record answers?	23.4	19.2	57.5
In which test was it easier to change answers?	5.4	4.2	90.4
Which test were you more likely to guess the answer in?	17.4	52.7	29.9
Which test was more comfortable to take?	37.7	12.0	50.3
In which test would you be more likely to receive the same score if you took it a second time?	31.1	41.9	26.9
Which test was more enjoyable to take?	12.0	12.0	76.0
Which test more accurately measured your reading comprehension skills?	50.9	26.3	22.8

Table 13 indicates that more than half of the participants either held already or 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 53% found it easier to read the text on paper than on screen. One interesting finding is the percentages of the last question about perception of the accuracy of the two modes for measuring the reading comprehension skill of the participants. About 51% of the subjects think that PBT measures their comprehension skills more accurately while only 23% think that CBT is better in this respect. I would argue that these percentages do not contradict their posterior testing mode preference where 34% chose PBT again while 52% went for CBT as their preferred testing mode.

We used one-sample t-test to examine the significance difference between the subjects' attitudes towards PBT and CBT features. Table 14 shows the t-test results:

Table 14: One-Sample t-test results

Questions	Test Value = 2	
	t	Sig.
In which test were reading passages easier to navigate through?	1.821	.070
In which test were reading passages easier to read?	-2.717	.007
In which test was the text in the items easier to read?	1.723	.087
Which test created less anxiety?	-1.155	.250
Which test was less fatiguing?	3.983	.001
In which test was it easier to record answers?	5.287	.001
In which test was it easier to change answers?	22.596	.001
Were you more likely to guess the answer in	2.396	.018
Which test was more comfortable to take?	1.743	.083
On which test would you be more likely to receive the same score if you took it a second time?	-.710	.479
Which test was more enjoyable to take?	12.045	.001
Which test more accurately measured your reading comprehension skills?	-4.472	.001

Difference in attitudes towards navigating through the reading passages, reading the text in the items, and test anxiety was non significant. Likewise, there was no significant difference in attitudes to comfort while taking the test and perception of receiving the same score. In contrast, (53%) liked reading passages on paper with a significant difference from those (33.5%) who liked reading the text on computer. Also (55%) liked the CBT because it is less fatiguing than the PBT. Moreover, (57%) and (90%) liked the CBT than the PBT, with a significant difference, because it is easier to record

and change the answers. A significant difference between the subjects who liked the CBT than those who liked the PBT on their attitudes towards which mode it is more likely to guess the answer in. (76%) liked significantly the CBT because it was more enjoyable than the PBT whereas (51%) liked the PBT more than the CBT because it seems a more accurate measure of their reading comprehension. On the whole, subjects seemed to have more significant positive attitudes to most of CBT features than PBT features. Yet, most subjects liked significantly the PBT than the CBT because reading the tests passages was easier on the PBT and it is a more accurate measure of their reading comprehension. These noticeable attitudes were justified in the retrospective interviews. Justifications varied from resisting change, accustomed to a specific testing mode, prior unpleasant experience with some computer-based course, to eye fatigue and unfamiliarity with new text presentation. These issues can be dealt with in different ways such as familiarising subjects more with CBT so subjects would be less change-resistant and more accustomed to the new testing mode. Furthermore, technology is rapidly developing; therefore, tackling issues like eye fatigue and text presentation is now achievable.

#### **4 Conclusion**

This study aimed to measure the comparability of paper- and computer-based L2 reading achievement tests as well as the construct validity of the new testing mode i.e. computer-based testing. It also investigated the relationship between several factors potentially affecting performance 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 almost no significant effect on the overall validity and reliability of the tests. We also reached a point where we can assert that computer familiarity has no influence on students' performance on computer-based tests. In addition, the other factors such as computer attitude and prior 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 findings of this study in relation to the test taking strategies used, however.

## References:

- Alderson, J. C. (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.
- Alderson, J. C. (2000) Technology in Testing: the Present and the Future. *System*, 28(4), 593-603.
- Boo, J. (1997) *Computerized versus paper-and-pencil assessment of educational development: Score comparability and examinee preferences*. Unpublished PhD dissertation, University of Iowa.
- Bresolin, M. J. (1984) *A comparative study of computer administration of the Minnesota Multiphasic Personality Inventory in an inpatient psychiatric setting*. Unpublished doctoral dissertation, Loyola University, Chicago, USA.
- Bugbee, A. (1996) The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education*, vol. 28-3, 282-300.
- Bunderson, V., Inouye, D. & Olsen, J. (1989) The four generations of computerized educational measurement. In R. L. Linn (Ed). *Educational Measurement*. pp.367-407. Phoenix, AZ: Oryx Press.
- Chalhoub-Deville, M. & Deville, C. (1999) Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics*, 19, 273-99.
- Chapelle, C. & Douglas, D. (2006) *Assessing language through computer technology*. Cambridge. UK: CUP.
- Chapelle, C. (2001) *Computer Applications in Second Language Acquisition: Foundations for teaching, Testing and Research*. CUP.
- Choi, I., Kim, K. and Boo, J. (2003) Comparability of a paper-based language test and a computer-based language test. *Language Testing*, vol. 20(3), 295-320.
- Christmas, O. (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*).
- Dalton, D. (1994) What do others know about CD-ROMs, LANs, modems, and more. A survey for staff training. *The Book Report*, 12, 19.
- Durndell, A. & Lightbody, P. (1993) Gender and computing: Change over time? *Computers in Education*, 21, 331-336.
- Eignor, D., Taylor, C., Kirsch, I. and Jamieson, I. (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.
- Fulcher, G. (1999) Computerizing an English language placement test. *ELT Journal*, 53(4), 289-299.
- Geissler, J. & Horridge, P. (1993). University students' computer knowledge and commitment to learning. *Journal of Research on Computing in Education*, 25, 347-365.
- Green, B., Bock, R., Humphreys, L., Linn, R. & Reckase, R. (1984) Technical guidelines for assessing computerised adaptive tests. *Journal of Educational Measurement*, 21, 374-359.
- Harrel, T., Honaker, M., Hetu, M. & Oberwager, J. (1987) Computerized versus traditional administration of the multidimensional aptitude battery-verbal scale: an examination of reliability and validity. *Computers in Human Behavior*, 3, 129-137.

- Hicks, M. (1989) The TOEFL computerized placement test: adaptive conventional measurement. (*ETS Reports No. 89-12*). Princeton, NJ: Educational Testing Services.
- Hofer, P. & Green, B. F. (1985) The challenge of competence and creativity in computerized psychological testing. *Journal of Counseling and Clinical Psychology*, 53, 826-838.
- Jegede, O. & Okebukola, P. (1992). Adopting technology in third world classrooms: students' viewpoint about computers in science teaching and learning. *Journal of Educational Technology Systems*, 20, 327-335.
- Johnson, N. & Green, S. (2004) *On-line assessment: the impact of mode on students performance*. A paper presented at the British Educational Research Association Annual Conference, Manchester, UK.
- Kirsch, I., Jamieson, J., Taylor, C. & Eignor, D. (1998) Computer familiarity among TOEFL examinees. *TOEFL Research Reports, .Report 59*. Princeton, NJ, USA: Educational Testing Services.
- Lee, J. (1986) The effect of mode of past computer experience on computerized aptitude performance. *Educational and Psychological Measurement*, 46, 727-733.
- Levin, T. & Gordon, C. (1989) Effect of gender and computer experience on attitudes towards computers. *Journal of Educational Research*, 5, 69-88.
- Loyd, B. & Gressard, C. (1984) The effect of sex, age, and computer experience on computer attitudes. *AEDS Journal*, 40, 67-77.
- McDonald, A. (2002) The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computing and Education*, 39, 299-312.
- Marcoulides, G. (1988) The relationship between computer anxiety and computer achievement. *Journal of Educational Computing Research*, 4, 151-158.
- Mazzeo, J. & Harvey, L. A. (1988) The equivalence of scores from automated and conventional education and psychological tests: a review of the literature. (*Report No. CBR 87-8, ETS RR 88-21*). Princeton, NJ: Educational Testing Services.
- Mazzeo, J. Druesne, B., Raffeld, P., Checketts, K. & Muhlstein, A. (1991) Comparability of computer and paper-and-pencil scores for two CLEP general examinations. (*College Board Report 91-5*). Princeton, NJ: ETS.
- Millar, F. & Varma, N. (1994) The effect of psychological factors on Indian children's attitudes toward computers. *Journal of Educational Computing Research*, 10, 223-238.
- Okinaka, P. (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*).
- Olsen, J., Maynes, D., Slawson, D. & Ho, K. (1989) Comparison of paper-administered, computer-administered and computerized achievement tests. *Journal of Educational Computing Research*, Vol.5, 311-326.
- Paek, P. (2005) Recent trends in comparability studies. *Pearson Educational Measurement Research Reports. Research Report 05-05*. Pearson Educational Measurement. USA.
- Parshall, C., Spray, J., Kalohn, J. & Davey, T. (2002) *Practical Considerations in Computer-Based Testing*. New York: Springer.

- Pomplun, M., Frey, S. & Becker, D. (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, 337-354.
- Powers, D. & O'Neill, K. (1993) Inexperienced and anxious computer users: coping with a computer-administered test of academic skills. *Educational Assessment*, 1, 153-173.
- Russell, M. & Haney, W. (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).
- Russell, M. (1999) *Testing on computers: A follow-up study comparing performance on computer and on paper*. Educational Policy Analysis Archive, vol. 7(20).
- Russell, M., Goldberg, A. & O'Conner, K. (2003) Computer-based testing and validity: a look back into the future. *Assessment in Education*, Vol. 10(3), 279-293.
- Sawaki, Y. (2001) *Comparability of conventional and computerized tests of reading in a second language*. *Language Learning & Technology*, vol.5 (2), 38-59.
- Stephen, D & Rowland, R. (1993) Initial IT training in Departments of Information and Library Studies in the British Isles: A survey of student views. *Education of Information*, 11, 189-204.
- Taylor, C., Jamieson, J., Eignor, D. & Kirsch, I. (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.
- Taylor, C., Kirsch, I., Eignor, D., Jamieson, J. (1999) Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning*, vol. 49-2, 219-274.
- Vincino, F. & Moreno, K. (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.
- Ward, T., Hooper, S. & Hannafin, K. (1989) The effects of computerized tests on the performance and attitudes of college students. *Journal of Educational Computing Research*, 5, 327-333.
- Weir, C. J. (2005) *Language Testing and Validation: An Evidence-Based Approach*. Palgrave Macmillan: NY, USA.
- Whitworth, B. (2001) *Equivalency of paper-and-pencil tests and computer-administered tests*. Unpublished PhD dissertation, University of North Texas.