COMPUTER-BASED VS. PAPER-BASED TESTING: ARE THEY THE SAME?

Saad Al-Amri

Computer-based vs. Paper-based Testing: Are They the Same?

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

Abstract

Comparability studies of paper-based tests and computer-based tests focused mainly on the equivalence of both forms and the contributing factors affecting this concept. There have been several studies in different fields such as education, psychology, ergonomics and L1 reading research (Sawaki, 2001). However, there has been no empirical study so far that has investigated the effect of prior computer familiarity on students' performance taking L2 tests (Chapelle & Douglas, 2006). Chapelle & Douglas (ibid) also mentioned the significance and lack of differential validity studies and how motivating it is to find out more about performance on computer-based testing. Sawaki (2001) argues that this type of empirical work should employ different methodologies such as eye movement, verbal protocols, post hoc interviews, and questionnaires to reveal useful results. Thus, this ongoing study examines the comparability of paper based tests and the computer based testing in L2 reading context, and the impact of test takers' characteristics, i.e., computer familiarity, computer preference, gender and test taking strategies on students' performance on computer based tests, and sequentially on comparability with paper based tests. 167 Saudi medical students took three reading achievement tests on both paper and computer. The Test of English as a Foreign Language (TOEFL) was devised to measure the students' proficiency and anchor the study tests. The study questionnaires focused on demographic information, participants' computer familiarity and preference of testing mode. The interview examined any change of preference after exposure to CBT. A triangulation of think aloud reports and post hoc interviews were employed to gain insight into strategies used on both testing modes, and to confirm comparability of both modes for validity purposes. The results are likely to reveal some information about the equivalence of both testing modes based on a scientific systematic perspective and have implications for the implementation of computer based reading tests in the context of the medical faculty EAP programme.

Introduction and Literature Review

Technological advancement has moved very rapidly since the last century. Computers became the most useful facilitator in achieving the majority of our goals. 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 influence of technology

on the essence of the assessment, which is the concept of validity. Moreover, little research has been conducted to investigate the interaction between the assessment mode and the test takers. There have been some studies that have focused on the comparability of the paper based testing and the computer based testing in some areas such as psychology, mathematics, and ergonomics (Sawaki, 2001). Yet, only a few studies have looked at this issue in the field of language assessment, such as those done by Taylor, et al (1999); Kirsch, et al (1998); Taylor, et al (1998); Eignor, et al (1998); Russell (1999); Russull 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, and neglected the process that resulted in these scores. Paek (2005) asserted the significance of computer familiarity, test taking strategies and academic subjects in measuring the equivalence of the two testing modes. Therefore, this study aims to measure the equivalence of the computer based and paper based testing, and consequently, the influence on the essence of the assessment, which is construct validity as defined by Messick (1989). This study also examines how test taker characteristics such as computer familiarity, preference, gender 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 construct validity. The methodology used in this study differs from the 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 conclusions of this study.

Evaluating the comparability and equivalence of both paper based and computer aided assessment is very crucial before introducing computer aided assessment into any context. It is always vital to ensure that both test qualities and test takers are not disadvantaged by shifting from one mode to another. It is also necessary to survey the readiness of the target context when deciding to implement the computer aided assessment. Conclusions and recommendations of this study will be of interest to the medical colleges in Saudi Arabia as they are the target audience in the target context.

Research Questions

This study attempts to answer the following questions:

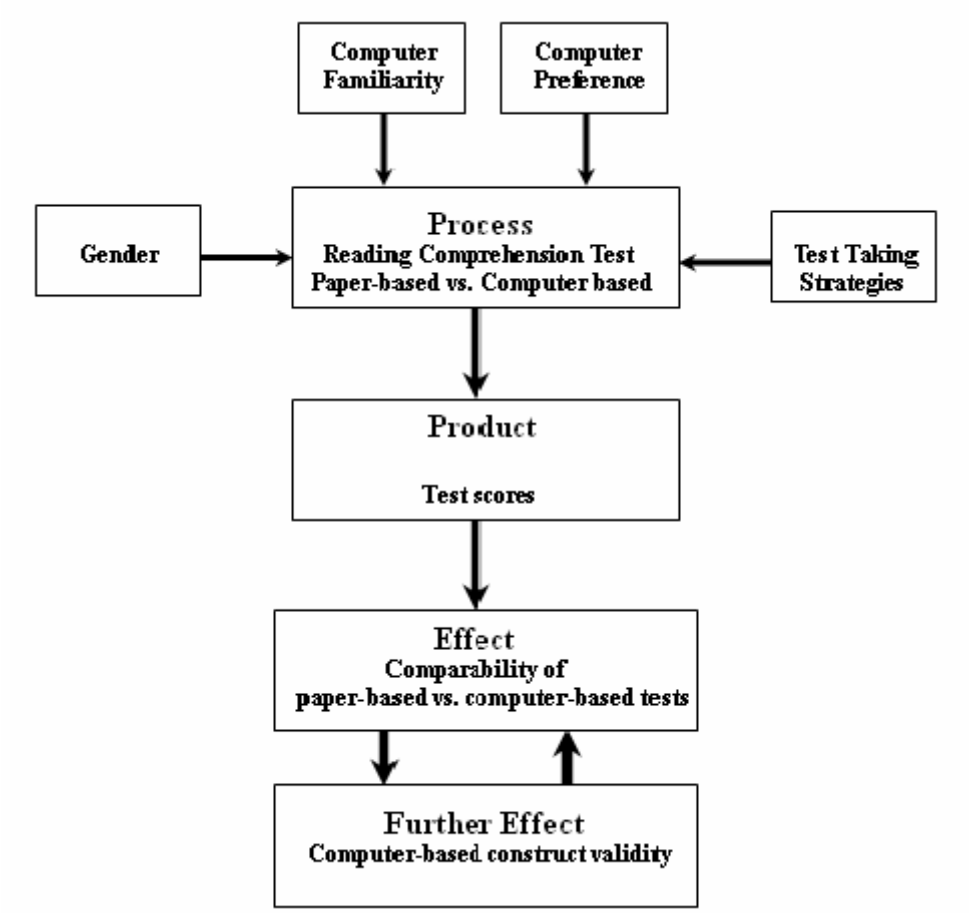
- Would construct validity be influenced by the test administration mode?
- Who will perform better on computer based tests: male or female and why?
- To what extent will students' performance change after exposure to a series of computer based tests?

- To what extent does computer familiarity affect students' performance on computer based tests?
- Do students use the same test taking strategies when taking the same test in two different testing modes?

Methodology

This study exploited the triangulation perspective in varying the sources of data collection. Thus, a range of quantitative and qualitative instruments were employed to gather the data required for this research. The following diagram explains the framework used in this study:

Study Framework



There were four instruments used to collect data for this project. They are as follows: the TOEFL test was used to measure the students' proficiency and to anchor the study tests. The study used two questionnaires that were built and designed based on questionnaires from the literature. The construction of the two questionnaires relied on adaptation of some existing questionnaires as well as implementing some new items to suit the study aims. Questionnaire one was designed to collect demographic data, measure computer familiarity

of participants, and computer preference of testing mode. The second questionnaire was used to collect data about preference after exposure to computer aided assessment. There were four computer assessment practical questions that were constructed to measure participants' computer familiarity in accordance with their responses on the first questionnaire. This aims at increasing the accuracy of measuring this construct. The study made use of 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. This tool measured score equivalence on both administration modes. Think aloud protocols were used to gain insight at the test taking strategies used by the participants when doing the same test on two different administration modes. Interviews were employed to collect more data about the participants' preference of CBT and the influence of shifting the testing mode on test taking strategies.

Data analysis

The data has been very recently collected and the coding and analyses have not been entirely finished. However, a very basic preliminary analysis has been conducted for the purpose of this paper. The full paper and analysis of both quantitative and qualitative data will be ready for publication at the 12th CAA conference.

First, the tests scores analysis showed that there is a difference in students' performance on paper based tests and computer based tests. This difference becomes quite obvious when looking at the mean scores of paper based and computer based tests. Table 1 shows a summary of all tests' mean scores and standard deviations.

Table 1. Descriptive Statistics

Tests	N	Mean	Std. Deviation
Score of paper test 1	167	77.45	12.879
Score of paper test 2	167	66.47	17.856
Score of paper test 3	167	67.90	18.362
Score of computer based test 1	167	74.65	15.389
Score of computer based test 2	167	61.89	19.022
Score of computer based test 3	167	64.07	15.259

Second, by running T-test, there is a high significant correlation between the means of the paper based and computer based tests.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Mean of Paper-based Tests	70.61	167	11.916	.922
	Mean of Computer-based Tests	66.87	167	13.215	1.023

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Mean of Paper based Tests & Mean of Computer based Tests	167	.738	.000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Individual Mean of Paper based Test - Individual Mean of Computer-based Tests	3.734	9.180	.710	2.332	5.137	5.257	166	.000

This difference cannot be interpreted immediately as the data concerning the independent variables, i.e., computer familiarity, gender, computer preference and test taking strategies, has not been coded and analysed yet.

Nonetheless, a simple calculation of those who have changed their preference before and after exposure to computer based tests can be noticed from these two questions. Question one was included in questionnaire one before exposure to computer based tests while question two was inserted in questionnaire two after the examinees had finished all the study tests; the following tables indicate the change in the participants' preferences:

(A question asked BEFORE exposure to CBTs)
Would you prefer taking tests?

		Frequency	Percent
Valid	On paper	68	40.7
	No difference	47	28.1
	On computer	52	31.1
	Total	167	100.0

(A question asked AFTER exposure to CBTs)
Which test do you prefer taking again?

		Frequency	Percent
Valid	On Paper	58	34.7
	No Difference	22	13.2
	On Computer	87	52.1
	Total	167	100.0

By comparing the frequencies of those who preferred paper based tests before and after taking computer based tests, we can see that 6% of those participants have changed their preference. Moreover, the percentage of those who do not mind taking the test in both modes dropped dramatically from 28.1% to 13.2%. On the other hand, this shift in preference resulted in an increase of 21% in the participants who favored the experience of computer based tests.

Further qualitative data was collected by interviewing 23 students sampled out of those participants who have changed their preference. These data have not been processed yet; however, we are certain that they will reveal more interesting findings.

Conclusion

This study aims to measure the comparability of both paper based and computer based L2 reading achievement tests. It also investigates the factors affecting the equivalence of both tests. Because the data coding and analyses are still in their early stages, no clear picture can yet be made about the final findings and results of this study. Also, most of the variables of this study cannot be discussed in this paper for the same reason. In addition, final solid conclusions cannot yet be drawn out of the available results due to the limited time available for analysis. This study is still ongoing and it is hoped that by the 12th CAA conference, the complete results and findings, as well as the entire project, will be ready for publication.

References

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

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

Chapelle, C. & Douglas, D., (2006). *Assessing language through computer technology*. Cambridge. UK: 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) pp. 295-320.

Eignor, D., Taylor, C., Kirsch, I. and Jamieson, J. (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.

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

Messick, S. (1989). *Validity in educational measurement* (3rd ed.). (ed) Robert Linn. London: Collier Macmillan Publishers.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effect for passage-based tests. *Journal of Technology, Learning and Assessment* Vol. 2 (6) pp. 1-44.

Russell, & 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).

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

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

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