ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol.3, Issue, 5, pp.170-176, May, 2011

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

RESEARCH ARTICLE

TO BE COMPUTERIZED OR NOT TO BE? A STUDY ON EFL LEARNERS' PERCEPTIONS TOWARDS VARIOUS MODES OF DELIVERY IN ENGLISH TESTS THROUGH ANALYTIC HIERARCHY PROCESS

Liwei Hsu

Department of Applied English, National Kaohsiung University of Hospitality and Tourism, 1, Sung-Ho Road, Hsiao-Kang District, Kaohsiung City, Taiwan

ARTICLE INFO

Article History:
Received 21st February, 2011
Received in revised form
25th March, 2011
Accepted 27th April, 2011
Published online 14th May 2011

Key Words:
Computerized test;
Mode of delivery;
Analytic Hierarchy Process (AHP)

ABSTRACT

This study aims to investigate EFL learners' viewpoints toward English test delivered in different modes; namely, computer-adaptive test (CAT), computer-based test (CBT) and conventional paper-pencil test (PPT). The participants were forty-nine (N=49) students from two colleges in Taiwan who had experience of taking all these three types of English tests. Additionally, five graduate students were invited to validate the contents of AHP questionnaire. Analytic Hierarchy Process (AHP) was the major research method employed by the present study to construct the framework of research. After a series of pair-wise comparisons processed with Expert Choice 2000 software package, weights of three objectives (convenience, fairness and computer experience) and six sub-objectives (physical limitation, immediate feedback, anxiety, accessibility, text presentation and response requirements respectively) were calculated to deduce the alternative. Among the three objectives, convenience ranked the highest followed by fairness and then familiarity. Subsequently, conclusion was drawn that CAT is the mode that best fitted EFL learners' expectation on English tests.

INTRODUCTION

The trend of globalization has made the status of English as the lingua franca more important than it was before. For English as foreign language (EFL) learners, being proficient in English is no longer a privilege but a basic requirement to get a job or be accepted by an academic program. The most persuasive way for a non-native speaker of English to prove his/her proficiency in English is through presenting the score of standardized tests like TOEIC or TOEFL. Modern technology has changed the landscape of language instruction as well as test delivery; therefore, computerized TOEFL (both Computer-based tests and internet-based tests) has gradually replaced paper-and-pencil based mode (PPT) by American colleges and universities to review applicants' English proficiency. Education Testing Service (ETS) has initiated the promotion of computerized TOEIC since the year of 2007. Regardless of the validity and reliability of question items, computerized testing has gained its momentum with advantages and disadvantages. The major advantages of computerized tests include their convenience, individualization and standardization; on the other hand, issues of fairness and economic efficiency are acknowledged as the disadvantages. Since most universities and colleges in Taiwan have set up a benchmark for graduates' English proficiency, it is noteworthy to acquire a comprehensive idea about Taiwanese college students' perception toward different delivery modes of English tests. However, only limited number of studies addressed the issue of potential effects of test-delivery-medium between conventional paper-and-pencil tests and computer-based tests (Chalhoud-Deville and Deville, 1999), particularly from examinees' perspective.

The Development of Computerized Language Tests

The very first time the term "computer" appeared in the language testing cycle can be traced all the way back to the year of 1935 when the IBM model of 805 was commercially available (Fulcher, 2000). The IBM 805 was initially designed to score multiple choice items to save labors on grading objective tests and it was soon adopted by Army language test administrators during the First World War. This modernized way of test grading had become very popular in the United States particularly in the era of the rapid expansion of school provision. The computerized scoring system has resulted to the prevalence of multiple choice items and this type of testing is still the bedrock of various measurements even in the present days. Computerized-assisted assessment (CAA) has revolutionized the way people think of the delivery of tests or assessments. Since the most dominant testing organization, ETS (Educational Testing Service), has adopted its

***Corresponding author:** liweihsu@mail.nkuht.edu.tw

computerized Graduate Record Examination (GRE) in October 1992 followed by completely computerized version of the Graduate Management Admissions Test (GMAT) in 1997, more and more tests developed by ETS started to apply this type of testing format (Wallace and Clariana, 2005). According to the study conducted by Conole and Bull in 2002, the tendency of replacing paper-and-pencil tests (PPT) by the CAA is prevailing and cannot be overlooked. Basically, there are two different kinds of computerized-assisted assessments; namely, the traditional computer-based test (CBT) and computerized-adaptive test (CAT). Wise and Plake (1989) defined CBT and CAT clearly from the perspectives of the test construct. CBT are the tests that use computer as a media to present the questions and collect examinees' answers instead of paper and pencil. Therefore, the constructs of CBT are similar to conventional tests. CAT is a relatively innovative way of constructing tests, which means that each test-take's items will not be exactly the same because the items will be selected based upon individual's previous responses (Glowacki; McFadden and Price, 1995). Because of this attribute, CAT is also acknowledged as individualized testing for being able to be tailored to individual differences (Wright and Stone, 1979).

Why Are Tests Needed to Be Computerized?

There is a list of advantages that computerized tests have over the paper-and-pencil tests (McNamara, 2000): first, examinees are able to receive the results of test immediately because the computer system can score items automatically. The second advantage is the efficiency and effectiveness of scoring for computerized tests (Spolsky, 1995, Jamieson, 2005). This "easy to be scored" attribute is also the main reason why multiple choice items has been so trendy in the past decades; nonetheless, the difference between "scorability" and "reliability" of a test shall be clarified. In other words, even though computerized tests are easier to grade, it does not explicitly endorse the reliability of such tests. In terms of advantages of the CATs, tests are able to be individualized to match test-taker's needs or the purpose of test being administrated. With a large amount of items in the test bank, it will be easier for the computerized test to identify examinees with extreme levels of proficiency. Moreover, level of difficulty of each item can be adjusted based upon each examinee's rate of correctness on all questions. The number of questions can be minimized to apprehend individual examinee's accurate level of proficiency; thus, they may feel much more convenient taking CAT than other tests.

This premise of large size of test bank cannot be overlooked either, especially when the results of test being analyzed through Item Response Theory (IRT). The rationale of IRT is beyond the scope of current study; however, valid interpretation on test scores is an essential issue for pertinent researches of testing. For this reason, constructing an item bank with sufficient number of questions with attributes of reliability, validity, impact, authenticity, interactiveness and practicality is the very first task for test designers and administrators to take into consideration (Fulcher, 2000; Cheng, 2006). Furthermore, there are also some advantages of using computerized language tests from human consideration perspective whatsoever (Brown, 1997). Such "humanistic" advantages include:

1. Examinees may feel more comfortable while taking computerized tests because they can work at their own pace.
2. Examinees do not need to spend too much time on a test; therefore, those who have physical limitations on staying at the same place for a long time will not be put through such "suffering."
3. Examinees may experience less frustration for taking computerized tests than paper-and-pencil ones because the items they are taking will be more close to their level of proficiency.
4. Examinees may have less mental pressure for taking computerized tests because there is only one question shown on the screen rather than the traditional ones with many questions on one page of booklet.
5. Previous studies have shown the fact that many students enjoy taking computerized tests (Steven and Gross, 1991).

Shumann (1997) has pointed out the effects of examinee's expectancy towards question items and how tests appeal to them do matter to their performance on a test. Specifically speaking, if a test-taker feels that he/she is going to succeed in a test, he/she may be more likely to have a positive attitude towards the test. By the same token, if the test-taker finds out the test is the most appealing to him/her, he/she tends to give it the best shot while taking it. These two points made by Shumann are important to test designers or administrators because the purpose of a language test is to elicit the examinee's performance in the target language which should reflect his/her competence, which also elucidate the importance of the present study.

The Issue of Fairness and Familiarity on English Tests

Another concern of researchers on computerized language tests will be the issue of fairness. Many prior researches have shown the fact that the change of delivery mode (from paper-and-pencil mode to computer-based mode) may change the level of difficulty of each item (Green, 1988; Sawaki, 2001). Moreover, the issue of test-taker's familiarity with computer use has come to lots of researcher's interest and concern. The results of some studies display a significant effect of individual examinee's computer familiarity on his/her performance (Lee, 1986; Buderson *et al*., 1989; Taylor *et al.*, 1999). Nevertheless, quite a few researches acquire different results, which indicate the fact that no such influence shall be significant enough toward test-takers' performance (Boo, 1997; Al-Amri, 2008). Case in point, whether computer familiarity is a major factor affects examinees' performance is still a controversial topic. Fairness of computerized tests also arise many researchers' interest. Kunnan (2000) specifically argued that accessibility (or affordability) to test equipments is the main concern of fairness toward any format of test. Therefore, in some countries or areas where no good electrical service or information systems are available to ordinary test-takers may cause the issue of fairness on computerized tests. A chance to rehearse on the format of a test or being able to attend test preparation courses does matter on the examinee's performance of a specific language test. Such effects of test-takers' being familiar with the test and comfortable with the test are described as face validity of a test (Brown, 2003). The statement posited by Guernsey (2008) revealed the unfair

situation to African students, which has made TOEFL being lack of face validity to examinees in Africa.

Comparability of Computerized Tests and PPT

In the academic communities, there are impressively numerous publications on the comparability between computerized tests and PPT (Wang and Shin, 2009). Some of these prior studies support the comparability of these two administration modes of tests (Kim and Hyunh, 2007; Paek, 2005; Wang *et al*., 2007, 2008; Kingston, 2009) whereas some findings claim to discover the differences between computerized tests and PPT with regard to examinees' performance in construed response items and selected response items (Neuman and Baysoun, 1998; McDonald, 2002; Choi and Tinker, 2002). Kingston (2009) proposed two possible reasons that caused such discrepancies: different administration systems used by different computerized tests and deficiency in research design, particularly the feasibility of assigning participants to control group and experimental group randomly. Even though he pointed out these two problems, the focus of his study rested on further analyzing the results of prior 81 studies on factors of examinee's grade and tested subjects through meta-analysis. What remains to be explored is the comparability of three administration modes (CAT, CBT and PPT) perceived by test-takers. In the present study, two previously mentioned downsides can be avoided through sample selecting process; additionally, the focal point is redirected from examinees' performance to their opinion about three forms of assessment.

Other Issues While Comparing Computerized Test and PPT

The definition of computer anxiety is coined by McDonald (2002) as "the fear experienced when interacting with a computer or anticipating an interaction" (p.305). There are impressively large quantities of publications on the influence of computer anxiety toward test-taker's performance on computerized tests (Wise *et al.*, 1989; Desai, 2001; Stricker, Wilder, and Rock, 2003; Smith and Caputi, 2007; Douglas and Hegelheimer, 2007). Intuitively, like other issues within the discipline of education, the actual consensus is difficult to attain. Regarding the source of computer anxiety, some scholars argue that an individual's computer experience and/or familiarity has a positive relationship with his/her level of anxiety while being placed in CBT settings (Thatcher and Perrewe, 2002; Hasan, 2003; Beckers and Schmidt, 2003; Broos, 2005), yet some findings claim that such association is not significant (Todman and Lawrenson, 1992; Durndell and Lightbody, 1994; Chua, Chen, and Wong, 1999). However, no cause-effect relationship has been settled from these studies. Besides, it is noteworthy that advocates of the influence of computer experience are newer than the other side chronologically. The development of modern technology makes computers more user friendly than a decade ago; therefore, the influence of computer experience is supposed to be lower. A further study is needed to address this question. By the same token, some scholars (Gos, 1996; Beckers and Schmidt, 2003) scrutinize this issue from another standpoint. The attention of their studies shifted from the quantity to quality of experience an individual have had with the computer. Therefore, this paper attempts to answer the call for a research that takes both quantity and quality of an individual's computer experience into account. Since participant's computer experience is another criteria proposed by the present study, which has received little attention in the academia (Sawaki, 2001), selecting appropriate sub-criteria crucial but difficult to the present study. Based on the arguments of Yu (2010) and Bennett (2003), the present study defines the effects that an examinee's computer experience toward a test as his/her acquaintance with text presentation and response requirements. The effect of text presentation onto Examinees' performance has been a popular topic for the comparability of computerized tests and conventional PPT. The empirical results yielded from prior research showed disagreements on this issue and thus expected more empirical evidences. It is to the author's knowledge that only limited amount of publications have tackled this issue from test-taker's perspective. Moreover, the response requirement is coined as the way examinees are supposed to input the answers to the question items (Yu, 2010) while text presentation means the layouts texts are designed to present on the paper or computer screen.

MATERIAL AND METHODS

Analytic Hierarchy Process (AHP)

Analytic Hierarchy Process (AHP) was developed by Saaty in 1971, which has been applied as a decision-making technique in various fields. Substantially, the AHP has been employed to obtain intensities of preference for the objectives and sub-objectives (in some studies, they are named as criteria and sub-criteria) by the different catchment stakeholder groups in multicriteria evaluation procedures for ranking alternative options. In the present study, the hierarchy is consisted of four layers; namely, the first level is the Goal, which can also be defined as the ultimate goal of decision-making process. The second level is the Objectives, which is constructed on the basis of the Goal and the third level is the sub-objectives with respect to objectives. The last layer is the Alternatives which refers to the choices for the participants. The implementation of AHP includes the following four steps:

1. Construct the relationship between levels.
2. Establish pair-wise comparison matrix.
   Compute weight values and Consistence Index (CI)/Consistence Ratio (CR) of individual levels.
3. Make priorities among objectives and sub-objectives and elicit the alternative that best fits the stakeholders' expectation.

The basic structure of AHP proposed by the present study is hereby presented as the following:

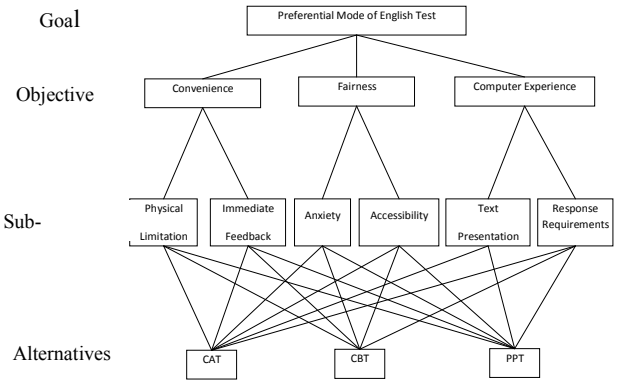

Fig. 1. The Proposed AHP Structure of the Present Study

Questionnaire Design

The design of AHP questionnaire is based on the pair-wise comparison; therefore, the traditional Likert 5-point scale is not appropriate for such design. Based on the reviewed literatures and the Goal of present study, the question items are developed accordingly. The architecture of the AHP was established on the advice of two professors of English to construct the interrelations of objectives and sub-objectives. Furthermore, in order to ensure the appropriateness of this questionnaire, five graduate students were invited to check the wordings and structures of question items. The final version of the questionnaire can be referred to the Appendix A.

Participants

In addition to the five graduate students as reviewers of the AHP questionnaire, fifty-three students from two colleges in southern Taiwan, who had experience of taking these three modes of tests, were invited to answer the AHP questionnaire. Due to the difficulty of recruiting subjects who have prerequisite experience of taking three types of English tests, total randomization is unrealistic in the present study. Instead, purposive sampling was adopted as a manner of the participant selection. After the questionnaire were collected in May, 2010 and subsequently examined the consistence, 49 validated questionnaires were used for computing the weight values through the Expert Choice 2000 software pack.

Data Analysis

As mentioned above, the retrieved data were scrutinized for the consistence before the AHP was performed. Data with consistence value higher than 0.2 would be partialed out by the present study. Pair-wise comparison matrixes between two items were formulated with the application of Expert Choice 2000. Participants' collective preference toward three modes of English tests was inductively yielded after a series of pair-wise comparisons were performed.

RESULTS

Demographic Information

According to the validated questionnaire, 23 participants were males whilst the other 26 were females. In terms of their academic background, most of them (around 90%) were English majors for their requirement of achieving a substantiate score of TOEIC, TOEFL (iBT, CBT or PPT) or IELTS as prove of their English proficiency. The rest of participants came from other hospitality programs such as Hotel Management (N=2) and Travel Management (N=3). The following Table 1 presents the demographic information of the participants.

Table 1. Demographic Information on the Participants

Gender	Number	Percentage
Males	23	46.94%
Female	26	53.06%
Academic Background		
Applied English	44	89.79%
Hotel Management	2	0.04%
Travel Managment	3	0.06%

Estimation of Relative Priorities of the Major Objectives

The average scores of the objectives and sub-objectives were computed with the Microsoft Excel. Afterwards, the Expert Choice software was administered to estimate the intensities of the importance of three major objectives (convenience, fairness, and computer experience) and their sub-objectives. The estimated weights represented the relative priorities of objectives and sub-objectives by the participants. Priorities for each objective responded by the participants, and inconsistency ratios for these, are presented in Table 2.

Table 2. Comparison of criteria with respect to the objectives

	Convenience	Fairness	Computer experience	Local Priority
Convenience	1	3	5	0.637
Fairness	1/3	1	3	0.258
Computer experience	1/5	1/3	1	0.105
λmax=3.039, CI=0.019, CR=0.033				
Inconsistency ratio=0.04				

Weights Assigned to Sub-Objectives

For each sub-objective at the lower level, pair-wise comparisons for each alternative had yielded participant's preference weights as reported in Table 3. For the Convenience aspect, "immediate feedback" is the feature that valued by the participants more than "physical limitation" when the convenience of a test is taken into account. In terms of the examinees' perception on the fairness toward test modes, the issue of anxiety was out-weighted by accessibility (0.333 to 0.667). Concerning the effect of their computer experience toward three different types of test, text presentation was posited to be more important than answer input (the weights were 0.833 and 0.167 respectively).

Table 3. The Weights of Sub-Objectives

Objectives	Sub-Objectives	Weights
Convenience	Physical limitations	0.200
	Immediate feedback	0.800
Fairness	Anxiety	0.333
	Accessibility	0.667
Computer experience	Text Presentation	0.833
	Response Requirements	0.167

Subsequently, the results of AHP derive the preferred mode of delivery among these tree tests for the participants. The synthesized ratings of three tests from high to low were CAT (0.562), CBT (0.221) and PPT (0.217). The CAT was considered as the most adequate test by the participants when all the proposed factors/issues were taken into consideration. The completed structure of AHP is depicted as the diagram below. Before the final decision was imminent, there is still one thing to do in the AHP, which is the sensitivity analysis.

Sensitivity Analysis

According to Alessio and Ashraf (2009), the last step of AHP is the sensitivity analysis, which is to response the query of the stability of the outcome to changes in the various factors in the hierarchy. Therefore, the main purpose of sensitivity analysis is to validate robustness and application of model. The sensitivity analysis of the present study indicated CAT indeed was valued greater by the participants in the convenience and fairness, but not the familiarity, which was dominated by

traditional PPT. The sensitivity analysis confirmed the results of AHP and asserted the status of computerized adaptive test as the most preferred alternative perceived by EFL learners. Result of sensitivity analysis is demonstrated in the Figure 3 below:

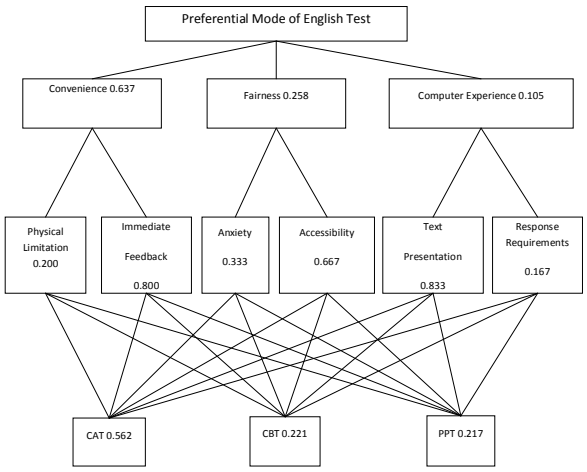

Fig. 2. The AHP Structure with Relative Weights

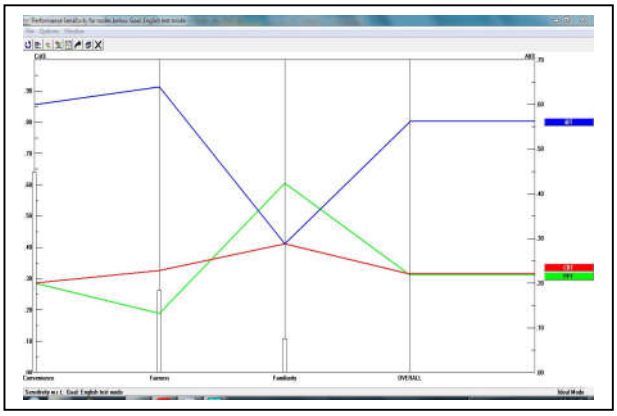

Fig. 3. Sensitivity Analysis of Three Criterion and Alternatives

DISCUSSION

This study attempted to investigate the examinees' viewpoints on three modes of delivery of English reading test. Three objectives (or criteria), along with six sub-objectives were included in the hierarchic structure based on the review of relevant literatures. Results of pair-wise comparisons among "Convenience," "Fairness" and "Computer experience" indicated that EFL learners considered the attribute of convenience a test could offer was the most important issue to them. Within the domain of convenience, being able to receive immediate feedback was significantly outweighed by the participants than the solution to physical limitation. Suchlike result is in line with previous studies on the importance of immediate feedback on their performance which can be offered by CATs (McNamara, 2000; Pino-Silva, 2008). Another explanation is that none of the participants was disable who may encounter one or some physical limitations while taking test. Yet, different story was depicted within the domain of the objective of fairness. Unlike the results derived from the prior research (Wise *et al.*, 1989; Desai, 2001; Stricker, Wilder and Rock, 2003; Smith and Caputi, 2007; Douglas and Hegelheimer, 2007), anxiety of taking computerized tests was counterintuitively not considered as a prime issue; instead, the accessibility to CAT or CBT was assessed to have greater priority to their final decision. The possible reasons led to this result are threefolds: first, the participants had taken too many English tests and they were no longer anxious toward taking English tests regardless of the mode of delivery. The other reason is that participants of the present study were from all walks of lives and the chance for them to rehearse different modes of English tests were unequal. These participants' anxiety can possibly be lowered by practicing if they had the opportunity to access the target test. The other explanation can be attributed to the feature of AHP to examine the relative priority between two variables for "it captures the spread of influence from the more important and general criteria to the less important ones (Saaty, 2006, p. 96)." In other words, the variable of anxiety was outweighed by accessibility not because of the unimportance of anxiety. We should interpret this result as the greater magnitude of accessibility.

Concerning the participants' computer experience and whether as such experience will affect their perception toward various modes of tests, the present study serves as one of the handful studies compare the importance of two variables—text presentation and response requirement. According to the result of AHP, text presentation was accounted for more magnitudes than response requirement. The way a test presents the text was found to cause various level of fatigue to test-takers (Dillon, 1992). Whether rehearsing on reading on the computer monitor can alleviate as such fatigue is another interesting topic to be investigated. The future study may conduct an experimental study on this issue. Moreover, participants posited that how the answers should be filled out on the answer sheet or key in through computer keyboard (and with mouse clicking and scrolling) are comparatively not that bothersome to them. They may feel that input answers on CAT or CBT is easier and more convenient to them because they do not have to circle the answers on the answer sheet which is time-consuming in some way. Consequently, participants of the present study showed their preference on the computerized test over the traditional PPT, which is in good agreement with the large body of previous studies (Brown, 1997; Steven and Gross, 1991; Kingston, 2009; Johnson and Green, 2004). Taking two modes of computerized test into further discussion, CAT significantly gained greater popularity by the participants for its largest marginal ratio. The major explanation is the convenience offered by the CAT, which can be attributed to the aforementioned advantage of immediate feedback. In terms of their perception toward CBT and PPT, both are identical for their convenience and fairness. The only difference between these two types of tests is the text presentation and response requirement which reflected on individual's computer experience.

Limitation

Inevitably, there are limitations to this study. The first major limitation arises with AHP known as 'rank reversal', which is associated with the relative nature of the judgments involved. If there is one more variable added to the structure, the results

may be changed dramatically. Second, the generalization of the results shall be cautiously used due to recruitment of participants with limited variety. Future studies are needed to be conducted in a large scale for more persuasive generalizability. Third, the selection of objectives and sub-objectives is solely on the basis of reviewed literatures. Most of previous works are conducted within the context dissimilar from the present study; thus, the application of Delphi technique to construct criteria with the collective consent of stakeholders is the solution toward this problem. The present study casts light on the direction for future research on the same topic.

Conclusions and Implications

In summary, we demonstrate that EFL learners' perception toward three modes of English tests through the performance of analytic hierarchical process. With the administration of a series of pair-wise comparison, the AHP indicates the convenience of a test has the greatest influence onto participants' choosing tests, followed by the fairness and then the familiarity. With respect to the convenience of a test, immediate feedback was considered as the one with greatest gravity. Accessibility to any type of test was perceived as much more important than anxiety when the issue of fairness is at stake. This is by far the most surprising and impactful finding elicited by the present study. Nevertheless, being familiar with the mode of delivery was not that decisive from participants' viewpoint. Even so, text presentation outweighed response requirements for acquiring higher score within this domain. Accordingly, CAT logically was the alternative best fitted all the conditions to participants' concern.

The findings of this study have some implications academically and practically. The academic implications are mainly based on the abovementioned research limitations encountered by the present study, which have been discussed above. Practically, the results of this study can inspire the designers or administrators of English tests redirect their focus on a test while they are preparing the next one. The innovativeness in convenience, particularly the immediate feedback, can be appealing to EFL learners. For classroom teachers who cannot afford CAT for every test administered in the class, providing test results to students as soon as possible is a compromising solution. Furthermore, the tests should be accessible to most, if not all, test takers for the stake of fairness. Finally, the layout of text presentation is also a critical issue taken by examinees. Even though most tests are delivered as PPT in class, the font size and line-width of the text cannot be overlooked for optimize the test implementation, especially within the EFL context where the examinees are assessed with a non-native language.

REFERENCE

Al-Amri, S. 2008. Computer-based testing vs. paper-based testing: a comprehensive approach to examining the comparability if testing modes. *Essex Graduate Student Papers in Language and Linguistics 10:* 22-44.

Alessio, I., and Ashraf, L. 2009. Analytic Hierarchy Process and Expert Choice: Benefits and limitations. *Insight, 22*(4): 201-220.

Beckers, J. J. and Schmidt, H. G. 2003. Computer experience and computer anxiety. *Computers in Human Behavior, 19*(6): 785-797.

Bennett, R. E. 2003. Online assessment and the comparability of score meaning. *ETS Research Report.* Retrieved on July 30, 2010 from http://www.ets.org/Media/ Research /pdf /RM-03-05-Bennet.pdf.

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

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

Broos, M. A. 2005. Gender and information and communication technologies (ICT) anxiety: Male self-assurance and female hesitation. *Cyber Psychology and Behavior, 8* (1): 145-166.

Brown, J. D. 1997. Computers in language testing: present research and some future directions. *Language Learning and Technology,* 1: 1, 44-59.

Chalhoub-Deville, M. and Deville, C. 1999. Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics, 19:* 273-299.

Cheng, Y. C. 2006. *The research on the application of computer assisted teaching in Chinese as a foreign language education.* Beijing: Commercial Publisher Co.

Choi, S. W., and Tinkler, T. 2002. Evaluating comparability of paper-and-pencil and computer-based assessment in a K-12 setting. *Paper presented at the 2002 annual meeting of the National Council on Measurement in Education.*

Chua, S. L., Chen, D. T., and Wong, A. F. L. 1999) Computer anxiety and its correlates: A meta-analysis. *Computers in Human Behaviors, 15*(5), 609-623.

Conole, G. and Bull, J. 2002. Pebbles in the Pond: Evaluation of the CAA, *Proceedings of the 6th Computerized-Assisted Assessment Conference,* Loughborough, 63-73.

Dillon, A. 1992. Reading from paper versus screens: A critical review of the empirical literature. *Ergonomics, 35*(10), 1297-1326.

Durndell, A. and Lightbody, P. 1994. Gender and computing 1986-1992: has any changed? In A. Adams, and J. Owen (Eds.), *Breaking old boundaries: building new forms. Proceedings of the 5th IFIP Conference on Women, Work and Computerisation.* Manchester: Manchester University Press.

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

Fulcher, G. 2000. Computers in language testing. In Brett, P. and G. Motteram (Eds.), *A special interest in computers* (pp. 93-107). Manchester: IATEFL Publications.

Glowacki, M. L.; McFadden, A. C. and Price, B. J. 1995. Developing Computerized Tests for Classroom Teachers: a Pilot Study, *Proceeding of the Annual Meeting of the Mid-South Educational Research Association,* Biloxi, 8-10.

Gos, M. W. 1996. Computer anxiety and computer experience: a new look at an old relationship. *The Clearing House, 69*(5): 271-276.

Green, B. F. 1988. Construct validity of computer-based tests. In H. Wainer and H. I. Braun (Eds.), *Test validity* (pp. 77-86). Hillsdale, NJ: Lawrence Erlbaum Associates.

Guernsey, L. 2008. Bowing to criticism, ETS suspends computerized tests in 20 African nations. Retrieved on 2008/9/16 from http://www.h-net.org/about/ press/articles /ets.html

Hasan, B. 2003. The influence of specific computer experiences on computer self-efficacy beliefs. *Computers in Human Behavior, 19*(4): 443-450.

Jamieson, J. 2005. Trends in computer-based second language assessment. *Annual Review of Applied L*inguistics, *25:*228-242.

Johnson, M., and Green, S. 2004. *On-line assessment: The impact of mode on students' strategies, perceptions, and behaviours.* Paper presented at the annual meeting of the British Educational Research Association, Manchester, Great Britain.

Kenyon D. M. and Malabonga, V. 2001. Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning and Technology, 5*, 2, 60-83.

Kim, D. H., and Hyunh, H. 2007. Comparability of computer and paper-and pencil versions of Algebra and Biology assessments. *Journal of Technology, Learning, and Assessment, 6*(4). Retrieved on December 13, 2009 from http://www.jtla,org.

Kingston, N. M. 2009. Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education, 22*(1): 22-37.

Kunnan, A. J. 2000. *Fairness and validation in language assessment.* Cambridge: Cambridge University Press.

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

MaNamara, T. 2000. *Lang*uage *Testing.* Oxford University Press. New York.

McDonald, A. S. 2002. The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers and Education, 39:* 229-312.

Neuman, G., and Baysoun, R. 1998. Computerization of paper-and-pencil tests: When are they equivalence? *Applied Psychological Measurement,* 22(1): 71-83.

Paek, P. 2005. Recent trends in comparability studies (PEM Research Report 05-05) Retrieved on December 20, 2009 from http://www.pearsonedmeasurement.com/downlods/ research/RR_05_05.pdf.

Pino-Silva, J. 2008. Student perceptions of computerized tests. *ELT Journal, 62*(2), 148-157. doi: 10.1093/elt/cclo56

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

Saaty, T. L. 2006. *Fundamentals of decision making and priority theory with the analytic hierarchy process.* Pittsburgh, PA: RWS Publications.

Shumann, J. H. 1997. *The Neurobiology of affect in language.* Malden, MA: Bloackwell Publisher.

Thatcher, J. B., and Perrewe, P. L. 2002. An empirical examination of individual traits as antecedents to computer anxiety and computer self-efficacy. *MIS Quarterly, 26*(4): 381-396.

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, 49*(2): 219-274.

Todman, J., and Lawrenson, H. 1992. Computer anxiety in primary schoolchildren and university students. *British Education Research Journal, 18*(1): 63-72.

Wang, H. and Shin, C. D. 2009. Computer-based and paper-pencil test comparability studies. *Test, Measurement and Research Service Bulletin of Pearson Education, 9:* 1-6.

Wang, S. Jiao, H., Young, M. J., Brooks, T. E., and Olson, J. 2007. A meta-analysis of testing mode effects in Grade K-12 mathematics tests. *Educational and Psychlogical Measurement, 67:* 219-238.

Wallace, P, and Clariana, R. B. 2005. Gender differences in computer-administered versus paper-based tests. *International Journal of Instructional Media, Vol. 32*

Wise S. L. and Plake, B. S. 1990. Computer-based testing in higher education. *Measurement and Evaluation in Counseling and Development, 23:* 3-10.

Yu, G. 2010. Effects of presentation mode and computer familiarity on summarization of extended tests. *Language Assessment Quarterly, 7*(2): 119-136.

****