*The Journal of Applied Linguistics Vol. 3, Issue 2*
*Fall 2010*

# Comparability of Computer-based and Paper-based Versions of Writing Section of PET in Iranian EFL Context

*Mohammad Mohammadi*[1]
*Masoud Barzgaran*
*Urmia University*

Computer technology has provided language testing experts with opportunity to develop computerized versions of traditional paper-based language tests. New generations of TOEFL and Cambridge IELTS, BULATS, KET, PET are good examples of computer-based language tests. Since this new method of testing introduces new factors into the realm of language assessment ( e.g. modes of test delivery, familiarity with computer, etc.),the question may be whether the two modes of computer- and paper-based tests comparably measure the same construct, and hence, the scores obtained from the two modes can be used interchangeably. Accordingly, the present study aimed to investigate the comparability of the paper- and computer-based versions of a writing test. The data for this study were collected from administering the writing section of a Cambridge Preliminary English Test (PET) to eighty Iranian intermediate EFL learners through the two modes of computer- and paper-based testing. Besides, a computer familiarity questionnaire was used to divide participants into two groups with high and low computer familiarity. The results of the independent samples t-test revealed that there was no statistically significant difference between the learners' computer- and paper-based writing scores. The results of the paired samples t-test showed no statistically significant difference between high- and low-computer-familiar groups on computer-based writing. The researchers concluded that the two modes comparably measured the same construct.

[1] Corresponding author. E-mail:

Keywords: Comparability Study, Computer-based Language Test (CBLT), Paper-based Language Test (PBLT)

The implementation of computer technology for language assessment dates back to the 1960s when computers were used for analysing test data or storing a large number of test items as test banks (Chapelle & Douglas, 2006). More recently, with the proliferation of personal computers, developing, modifying, and even administering language tests have become more practical. Yet, some advantages like immediate scoring and reporting of results, opportunity to include innovative item formats, and reduced costs of test production, administration, and scoring, add more and more to the popularity of the computer-based tests (CBTs) over the traditional paper-based tests (PBTs) (Wang& Shin, 2009).

Essay writing parts of standardized tests, however, show little flexibility with the process of computerizing traditional paper-based tests, at least in terms of scoring. That is, with current technology, still human raters are needed to score word-processed essays. More importantly, computer-based mode of test delivery and test taking is likely to impose considerations regarding the writing performance of test takers per se and accordingly the validity of such tests.

The present study aims to deal with the comparability issues related to computer- and paper-based writing assessment based on the following considerations that the researcher has traced in this review of the related literature.

Firstly, since comparability of CBTs and PBTs is a multifarious issue, numerous factors and variations must be taken into account in studying the comparability of a computer-based test and its traditional paper-based counterpart. These variations include content areas, participants' familiarity with computer, data collection design, and item format (Wang & Shin, 2009). During the last two decades, numerous studies have been conducted on the comparability of computer-based language tests (CBLTs) and paper-based language tests (PBLTs). However, most of these studies were conducted in the 1980s and early 1990s, when the

current word processors were not present or in large-scale use. In addition, the students who participated in these early studies were generally less familiar with computer technology compared to students today (Goldberg, Russell, & Cook, 2003). Nevertheless, it seems that students' skills related to typing and working with word processors progress at a fast pace compared with their other computer-related expertise. This is, however, a disputable issue regarding the learners of English as a Foreign Language whose mother tongue has a different orthographical system from English.

Secondly, multiple-choice tests seem to be the general interest of the so far conducted research. Hence, little attention has been paid to open-ended tests such as writing assessment. Open-ended test tasks appear to be more prone to the impact of computer than other types. For example, when a multiple-choice grammar test is adapted to a computer-based version, there is only a shift from marking or circling a word on the paper to clicking or checking in a box on computer. However, in the case of open-ended tests and particularly essay writing, the whole story changes and new considerations emerge by shifting the medium of test taking from paper to computer. The very process of writing, written products, and even the scoring process of this test task are prone to the impact of computer and scant attention to these considerations are very much likely to lead to various problems on the way of validating standardized tests.

Issues pertinent to construct validity are of utmost importance in validating a standardized language test. Therefore, it is mandatory to ensure that the computerized and the conventional paper-based versions of a standardized language test equivalently measure the same construct(Chapelle & Douglas, 2006). More specifically as is the interest of the present study, deciding on the comparability of CBLTs and PBLTs in measuring writing ability of language learners demands more empirical research than those conducted so far.

To date, several studies conducted on this issue have come up ironically with rather inconsistent conclusions (Choi, Kim & Boo, 2003). To overcome this inconsistency in the findings of relevant studies, two resolutions have been recommended by the

researchers of language assessment.(Chapelle & Douglas, 2006)Primarily, holistic comparability studies need to be narrowed down to item-level studies. Moreover, numerous up-to-date research studies that can be conducted in various local settings and ESL/EFL contexts are likely to be helpful in alleviating the existing discrepancy. These considerations underscore the significance of the present study as it puts an emphasis on the item-level comparability study and focuses on only one construct (writing ability) in a new EFL context (Iranian EFL context).

To investigate the comparability of the written products of intermediate EFL learners across the two modes of computer- and paper-based testing of writing the following questions are proposed:

Q1: Is there a significant difference between the writing scores of Iranian EFL learners' essays across computer- and paper-based writing tests?

Q2: Is there a significant difference between the writing scores of learners with high computer familiarity and low computer familiarity on computer-based writing test?

## Literature Review

Researchers over the past 20 years, through a number of cumulative researches, have introduced some areas of concern in the validation of computer-based language tests (CBLT). Chapelle and Douglas (2006) identified six potential threats to the validity of CBLT as a synthesis of the concerns expressed by various researchers. These potential threats are:

- different test performance
- new task types
- limitations due to adaptive item selection
- inaccurate automatic response scoring
- compromised security, and
- negative consequences. (p. 42)

One of the most common concerns about the validity of computer-based language tests is that there is a probability that test takers perform differently on CBLTs simply as a result of change

in the mode of test delivery. This means that a given CBLT may reflect ability or abilities different from those measured by a corresponding paper-based version. Obviously, as Chapelle and Douglas state, "it is a threat only to the extent that score users intend the two scores to be equivalent" (p. 42). As most of current large-scale standardized language tests are administered concurrently in both computer- and paper-based modes, offering examinees the opportunity to choose either of the modes, the two sets of scores obtained from the two modes need to be equivalent such that the users of these scores infer that they are indicators of the same abilities, and that the test is valid and enjoys an adequate degree of the characteristics of validity.

Chapelle and Douglas (2006) suggest that the threat of different test performance can be dealt with through: (1) interpretation of computer-method effect, and (2) test comparison study, in which the performance of test takers are compared on two tests which are the same except for the mode of test delivery, i.e, one form of the test is delivered as a computer-based test and the other as a paper-based one. The present study, thus, falls in the category of comparison studies which investigates whether there is a meaningful difference between the performances of L2 learners across the two modes of a language test. However, the study is specifically concentrated on the writing ability to find out if a computer-based writing assessment measures the same ability as that measured by a conventional paper-and-pencil test.

Presumably, one may probably ask, 'How can a comparability study address issues related to construct validity?' Construct validation uses theory or logic to develop hypotheses about the correctness in measuring the construct it claims to measure. However, in the case of comparability studies, as Lottridge, Nicewander, Schulz, &Mitzel (2008) point out, the construct validation paradigm is simplified to some extent because in a comparability study the nature of the construct being measured by two tests (or two testing modes) does not have to be identified. Rather, the researcher seeks to find out whether the constructs assessed by the two tests are the same.

So far, a number of research studies have been dedicated to the comparability issues of CBLTs and PBLTs without being unanimously conclusive about the comparability between them. Mazzeo and Harvey (1988 cited in Wang & Shin, 2009) provided one of the earliest reviews of the research on this topic and included around 30 comparability studies. Revealing mixed evidence regarding the comparability of CBLTs and PBLTs, their review came up with the conclusion that the test mode seemingly had no effect on power tests, but a considerable effect on speed tests. Their review also indicated that CBLTs tended to be more difficult than the PBLT versions. Kim (1999) performing a meta-analysis of ability measure tests found CBLTs and PBLTs as having comparable average scores (Wang & Shin, 2009). In a similar study, Kingston (2009 cited in Wang & Shin, 2009), synthesizing 81 comparability studies in K-12 multiple-choice tests which had been conducted between 1997 and 2007 found that the estimated effect size across all the studies was small. Most of the researches, however, provide ambiguous or conflicting findings mainly because of idiosyncratic differences in many variables, including previous exposure to computers, attitudes toward computers, intelligence and educational background (Mazzeo& Harvey, 1988; Mead &Drasgow, 1993; Schaeffer et al., 1993; Russel& Haney, 1997; Vispoelet al., 1997; 2001 all cited in Choi, Kim & Boo, 2003).

The literature relevant to the exclusive comparison of computer and paper-based essay writing performance is mixed in a manner similar to that of the general research on the comparability of CBLTs and PBLTs.While some studies have reported higher performance on essays written on computer in comparison to handwritten essays (Russel&Plati, 2001; Russel& Haney 1997 cited in Way, Davis & Strain-Seymour, 2008), some other studies (Way & Fitzpatrick, 2006; Bridgman & cooper, 1998 cited in Way et al., 2008) have come up with contrary findings suggesting a lower performance for computer-based essays compared with handwritten ones. A number of studies, on the other hand, have found no significant difference between compositions across the two modes.Collier and Werier (1995 cited in Lee, 2004) found that

although habitual computer writers were discomforted with paper writing, their performance on paper and computer-based composition was similar. Similarly, Young-Ju Lee (2005), studying six Korean ESL students' writing performance tried to find if there was a plausible difference in composing processes when they write timed-essay tests on paper and on computer. Although the number of subjects was too small to draw a generalization, the results of the study suggested that the difference between the essay scores across the modes was not significant.

The considerations pointed out above provided the researchers of this study with substantiated rationale for concentrating on writing assessment through computer. The present study was carried out as a quasi-experimental comparison study to explore possible differences that might exist between Iranian EFL learners' writing performance across the two modes of computer- and paper-based writing assessment.

## Method

### *Participants*

The participants of this study were selected from among the learners of a private language institute (Shokooh Language Institute) in Salmas, West Azerbaijan. The total number of learners who were assigned by the institute to intermediate level was 103. At the beginning of the study, the group of 103 intermediate learners was divided through a survey into two groups of learners with high and low computer familiarity. Later, the learners' homogeneity in terms of language proficiency was confirmed with the aid of a paper-and-pencil test and ultimately 80 students (N=80) were selected from the two computer familiarity groups as the sample of the study. The age of the learners ranged from 15 to 20, and the mother tongue of most of them was Turkish.

### *Instruments*

Cambridge Preliminary English Test (PET) and a standardized questionnaire of computer familiarity designed and

validated by the Organization for Economic Co-operation and Development, Program for International Student Assessment (OECD PISA) served as the two instruments of the present study. The questionnaire of computer familiarity was used prior to the data collection phase in order to determine the level of the learners' computer familiarity. As for the PET test, the listening, reading and writing sections of the paper-based version and only the writing component of the computer-based version were used in this experiment. The listening and reading parts of the paper-based PET that comprised items pertinent to structure, vocabulary, and reading comprehension yielded the required information to decide on the homogeneity of the learners. In the data collection phase of the experiment, however, only the writing subsections of both paper- and computer-based PET were used.

*Design*

A repeated-measures design was employed in the experiment since two measures of the learners' writing ability had to be compared with each other. The participants' writings, both on computer and paper, were measured through two separate tests with comparable, though slightly different, prompts. Slightly altered prompts were used to eliminate the possibility of practice effect, which, as a threat to test reliability, enables test takers to take advantage of their previous test taking experiences. However, the prompts had been chosen as comparable as possible so that they elicit similar schema or background knowledge on the part of learners, and that their difference would not affect the results of the study.

At the phase of writing assessment, the participants were randomly assigned to two groups of A and B. Each group comprised equal number of high and low computer-familiar learners (20 of each). Group A wrote on prompt 1 in the computer mode first and prompt 2 in the paper-and-pencil mode. In contrast, group B wrote on prompt 1 in the paper-and-pencil mode first and prompt 2 in the computer mode. This counterbalanced design was implemented to neutralize the sequence effect of the two tests and

the interaction of the prompts across the modes. Table 1 illustrates the design of this study in summary:

Table 1.

*The Order of Test Taking Across the Two Modes*

| Group | prompt 1 | prompt2 |
|---|---|---|
| A | Computer | paper-and-pencil |
| B | paper-and-pencil | Computer |

As for scoring, two raters scored the essays independently in order to increase the reliability of the test scores. The correlation of the scores obtained from the two raters was analyzed to determine the inter-rater reliability. Each essay was independently rated by the two raters based on the holistic grading benchmarks developed by ESOL examinations department of Cambridge University for scoring the writing section of PET

## Results

### *Results of Proficiency Pre-test*

The participant's proficiency test scores were grouped according to the computer familiarity group that they belonged to. Owing to the fact that an independent-samples t-test is sensitive to outlier scores, all the outlier scores from the two groups were identified with the aid of the SPSS software and excluded from the sample. The two sets of scores were subjected to an independent-samples t-test, the results of which are shown in Tables 2 and 3. With a glance at Table 2, one can infer that the scores of the two groups of high computer familiarity with the mean value of 45.15 (SD = 4.42, N= 52) and low computer familiarity with the mean value of 44.9 (SD = 3.83, N=44) are close to each other.

Table 2.
*High- and Low-computer-familiar Groups' Mean Scores on the Proficiency Pre-test*

| | group | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| proficiency score | high familiarity | 52 | 45.1538 | 4.42535 | .61369 |
| | low familiarity | 44 | 44.9091 | 3.83871 | .57871 |

Table 3.
*Independent Samples T-Test Verifying the Homogeneity of the Participants in Terms of Language Proficiency*

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 95% Confidence Interval of the Difference | |
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | Lower | Upper |
| proficiency score | Equal variances assumed | .799 | .374 | .287 | 94 | .775 | .24476 | .85361 | -1.45010 | 1.93961 |
| | Equal variances not assumed | | | .290 | 93.934 | .772 | .24476 | .84351 | -1.43007 | 1.91958 |

However, the data in Table 3 provide a more precise evidence of the homogeneity of the two groups. From one hand, the results of Levene's test show that the variances related to the scores of the two groups are equal because the p-value of the variances is greater than $\alpha = 0.05$ (p-value = 0.374 > 0.05 = α). On the other hand, the p-value related to the equality of means is greater than $\alpha = 0.05$, which indicates that the difference between the means of the two groups is not statistically significant (two-

tailed p-value = 0.775 > 0.05 = α), and hence the two groups of the participants with high and low computer familiarity are homogeneous in terms of language proficiency.

From each of the two groups of computer familiarity equal numbers of 40 participants were randomly selected to compose the final experimental sample (N = 80) whose writing scores on the two mediums of computer and paper were analyzed to test the hypotheses proposed by the researchers.

*Inter-rater Reliability*

All the participants' essays, written on computer and paper, were scored twice by two raters in order to determine the reliability of the scores. The relationship between the scores given by the two raters to computer- and paper-based essays was investigated using Pearson product-moment correlation coefficient (Tables 4 and 5).

Table 4.
*Correlation Between the Scores of Computer-based Essays Given by the Two Raters*

| | | rater 1 CBTwriting | rater 2 CBTwriting |
|---|---|---|---|
| rater 1 CBTwriting | Pearson Correlation | 1 | .897** |
| | Sig. (2-tailed) | | .000 |
| | N | 80 | 80 |
| rater 2 CBTwriting | Pearson Correlation | .897** | 1 |
| | Sig. (2-tailed) | .000 | |
| | N | 80 | 80 |

**. Correlation is significant at the 0.01 level (2-tailed).

Table 4, (r = +0.89, n = 80, p = .000 < 0.01) verifies a strong correlation between the scores of the raters in computer format. This helps to ensure the reliability estimates of the scores of both raters in scoring computer-based essays.

As for the scores of the paper-based essays, the results of Pearson product-moment correlation coefficient revealed that there was also a strong positive correlation ($r = +0.85$, $n = 80$, $p = .000 < .01$) between the scores of the two raters in the paper format.

Table 5.
*Correlation Between Scores of Paper-based Essays Given by the Two Raters*

| | | rater 1 PBTwriting | rater 2 PBTwriting |
|---|---|---|---|
| rater 1 PBT writing | Pearson Correlation | 1 | .859** |
| | Sig. (2-tailed) | | .000 |
| | N | 80 | 80 |
| rater 2 PBT writing | Pearson Correlation | .859** | 1 |
| | Sig. (2-tailed) | .000 | |
| | N | 80 | 80 |

**. Correlation is significant at the 0.01 level (2-tailed).

After averaging the scores of the two raters in each mode, a single score was given to the writing performance of every participant in either of the formats. The rest of the analysis including testing the research hypotheses was carried out based on these averaged scores. However, since t-test procedure of data analysis had to be employed, the data primarily needed to be normally distributed in order to meet the assumptions of parametric statistics. Accordingly, a one-sample Kolmogorov-Smirnov test and the histogram of the distribution of the scores of each mode were used to investigate the normality of the scores.

The *p*-value of the Kolmogorov-Smirnov test for all computer-based writing scores was greater than 0.05, which meant that the test distribution is normal (Asymp. Sig. (2-tailed) = p-value = $.206 > .05 = \alpha$). Similarly, the results of the Kolmogorov-

Smirnov test for all paper-based writing scores verified the normality of the distribution of these. The *p-value* of the Kolmogorov-Smirnov test was greater than the alpha level (Asymp. Sig. (2-tailed)= *p-value* = 0.40 > .05 = α), which denoted the normality of the paper-based writing scores.

*Results of analysis for Research Question 1*

Based on the first research question, regardless of the level of computer familiarity, each participant's writing score on computer was compared with his/her score on paper in order to investigate if their writing quality changed across the two modes. In other words, it was investigated that whether the mode of test delivery as an independent variable could act as a source of difference in the participants' writing scores. For this purpose, a paired-samples t-test was used since two measures of each participant's writing ability had to be compared with each other.

The results of the paired-samples t-test have been illustrated in Tables 6 and 7. As Table 6 shows, the mean value of the computer-based scores with N = 80 and SD = 2.20 is 11.325, while the mean value of the paper-based scores with N = 80 and SD = 2.185 is 11.343. Although the approximation of the two mean values is perceivable at first sight, Table 7 provides more precise information.

Table 6.
*Paired Samples Statistics*

| | | Mean | N | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| Pair 1 | total computer scores | 11.3250 | 80 | 2.20342 | .24635 |
| | total paper scores | 11.3438 | 80 | 2.18539 | .24433 |

Table 7
*Paired Samples Test Comparing the Two Sets of Computer- and Paper-based Scores*

| | | Paired Differences | | | | | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference | | | | |
| | | | | | Lower | Upper | | | |
| Pair 1 | total computer scores - total paper scores | -.01875 | .86216 | .09639 | -.21061 | .17311 | -.195 | 79 | .846 |

Since the alpha level has been set on 0.05 and the *p-value* of the test is greater than α (two -tailed p-value = .084 > .05 = α), it is concluded that the difference between the two sets of scores is not statistically significant. Furthermore, the fact that the confidence interval of the difference (lower = -.2106, upper = .1731) includes zero, and that the absolute value of *t* is less than 2, substantiates the researchers' deduction.

Therefore, based on the results of the paired-samples t-test, it was verified that the mode of test delivery did not result in significant difference between the language learners' writing quality on both modes of computer- and paper-based writing assessment.

According to Lottridge et al. (2008), if the two modes are comparable in terms of overall scoring of a sample, it can be reasonably assumed that the constructs measured by the two modes are the same. The analyses conducted for the first research questions revealed that the two modes are comparable in terms of overall scoring of the sample, which provides tenable evidence to conclude that the constructs measured by the two modes are the same. However, to add more credence to this assumption, the researchers investigated the go-togetherness of the participants' writing ability measures obtained from the two modes by using

Pearson product-moment correlation coefficient. Table 8 illustrates the result of the correlation analysis.

Table 8.
*Pearson Product-moment Correlation Between the Scores of Writing on Paper and Computer*

| | | total computer scores | total paper scores |
|---|---|---|---|
| total computer scores | Pearson Correlation | 1 | .923** |
| | Sig. (2-tailed) | | .000 |
| | N | 80 | 80 |
| total paper scores | Pearson Correlation | .923** | 1 |
| | Sig. (2-tailed) | .000 | |
| | N | 80 | 80 |

**. Correlation is significant at the 0.01 level (2-tailed).

As Table 8 indicates, there is a strong positive correlation(r = +.923, N = 80, sig.(2-tailed) = .000) between the paper-based writing scores and the computer-based ones. The result of the correlation analysis corroborates the researchers' conclusion made based on the findings related to the first and second research questions.

In general, it can be concluded that the two modes of computer-based and paper-based testing of writing comparably measure the same construct.

*Results of Analysis for Research Question 2*

In the second research question the participants' level of computer familiarity was taken as an independent variable to investigate if a test taker's level of computer familiarity, as a construct irrelevant ability, can affect his/her performance in computer-based writing assessment. For this purpose, an independent-samples t-test was utilized to compare the computer-based writing scores of the two groups of learners with high and low computer familiarity. Tables 9 and 10 show the results of the independent samples t-test.

As Table 9suggests, the mean score of the high-computer-familiar group N = 40, SD = 2.3,M = 11.50 seems to be slightly greater than the mean score of the low-computer-familiar group N = 40, SD = 2.09, and M = 11.15. However, the significance of this difference can only be determined by interpreting the data in Table 10. According to Levene's test in Table 10, the difference between the variances of the two groups is not statistically significant. Moreover, the t-test for equality of means with df = 78 ,t = 0.70, and a p-value (sig. = 0.76) greater than alpha level (0.05) indicates that there is not a statistically significant difference between the scores of the two groups. Furthermore, according to Table the 10 confidence interval of difference (lower = -.634, upper = 1.334) includes zero, which corroboratively reinforces the conclusion that the difference is statistically insignificant.

Thus, the results of the analysis imply that the difference between the writing scores of the participants with high computer familiarity and that of those with low computer familiarity was statistically insignificant. This means that the level of computer familiarity does not act as a source of construct-irrelevant variance in computer-based writing assessment.

Table 9.
*High and Low-computer-familiar Groups' Mean Scores on the Computer-based Writing Test*

| | group | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| total computer scores | high familiar | 40 | 11.5000 | 2.31771 | .36646 |
| | low familiar | 40 | 11.1500 | 2.09762 | .33166 |

Table 10.
*Independent Samples T-Test Comparing the Scores of High-and Low-computer-familiar Groups on the Computer-based Writing Test*

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 95% Confidence Interval of the Difference | |
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | Lower | Upper |
| total computer scores | Equal variances assumed | .090 | .765 | .708 | 78 | .481 | .35000 | .49426 | -63400 | 1.33400 |
| | Equal variances not assumed | | | .708 | 77.236 | .481 | .35000 | .49426 | -63415 | 1.33415 |

## Discussion

The purpose of the present study was to evaluate the comparability of computer-based and paper-based versions of Cambridge Preliminary English Test (PET) in measuring writing ability of Iranian intermediate EFL learners. This study was important, as pointed out by earlier research (e.g., Choi, Kim& Boo,2003; McDonald, 2002), to determine if administering tests in computer- and paper-based formats affect the comparability of scores obtained from these two testing formats. Moreover, it has been suggested that the mode of test delivery per se, as well as the mode -inherent characteristics on the part of test takers and raters may introduce construct-irrelevant variances into the scores.

The present study was conducted in accordance with some guidelines that have been published for examining comparability between CBTs and PBTs. For example, American Psychological Association (1986 cited in Lottridge et al., 2008) and the International Test Commission (2005cited in Lottridge et al., op.cit) provide the following guidelines:

> Scores from conventional and computer administrations may be considered equivalent when(a)rank orders of scores of individuals tested in alternative modes closely approximate each other, and (b) the means, dispersions, and shapes of the score distributions are approximately the same, or have been made approximately the same by rescaling the scores from the computer mode. (APA, 1986, p. 18)

> Provide clear documented evidence of the equivalence between the CBT/Internet test and non-computer versions (if the CBT/Internet version is a parallel form). Specifically, to show that the two versions: have comparable reliabilities; correlate with each other at the expected level from the reliabilities; correlate comparably with other tests and external criteria; and, produce comparable means and standard deviations or have been appropriately calibrated to render comparable scores. (ITC, 2005, p. 21)

The results of the analyses indicate that the findings are likely in line with those of Lottridge et al. (2008). They suggest that the comparability of two testing formats in terms of overall scoring of a sample reasonably entails construct equivalence. This reasoning is grounded in the assumption that "there is no counter evidence that differing constructs are involved when the score distributions are comparable" (pp.1-2).

In the present study, the comparability of CBT and PBT versions of a writing test was evaluated. The findings of the study suggested that there was no statistically significant difference between language learners' writing performance across the two modes. Furthermore, the results of the analyses revealed that the difference between the writing scores of high- and low-computer-familiar groups was not statistically significant, though it might intuitively be expected that learners with high computer familiarity would outperform low familiarity group in computer-based writing. One possible justification for this finding may be that the participants labelled as low-computer-familiar group were not completely unfamiliar with computer. In fact, as it actually seems, a threshold level of computer familiarity might have equipped the learners with sufficient hands-on and cognitive skill to write their essays on computer more or less conveniently.

As is evident in general, the findings of this study are positive and suggest that the computer-based and paper-based versions of essay writing section of PET are comparable in terms of overall scoring of the sample and measuring the same constructs. The findings of the study, thus, back up the conclusions of some other researchers who supported the comparability of CBT and PBT formats. For example, regarding the effect of computer familiarity Fulcher (1999), and Taylor, Kirsch, Eignor, and Jamieson (1998) found that computer familiarity or preference for either mode had no significant effect on students' scores. In a similar way as with the present study, the comparability studies conducted by Harrington, Shermis, and Rollins, (2000), Choi, Kim, and Boo (2003), H. K. Lee (2004),Puhan, Boughton, and Kim, (2007),Wang, Jiao, Young, Brooks, and Olson (2008), and Kingston (2009 cited in Wang & Shin, 2009) summed up with evidence supporting the comparability of CBT and PBT.

Consequently, as Lottridge et al. (2008) suggest, the comparability of computer-based and paper-based modes of testing is, ultimately, a matter of judgment. More clearly, the investigator's interpretation of the results of statistical estimates involves human judgment, probabilistic reasoning, and the strengths and limitations of the study design. For example, it is rather a rule of thumb to

decide on the comparability of the constructs measured by the two modes based on the score distributions, standard deviations, and correlations. As is well known, the construct validity issues can best be tackled through a sophisticated procedure of factor analysis. Thus, the interpretation of evidence is of great significance to the decisions regarding comparability of the two testing modes.

## Conclusion

The present study investigated the comparability between writing assessment through computer technology and traditional paper-and-pencil mode. The writing scores of 80 Iranian intermediate EFL learners on the two modes were subject to statistical analysis. The results of the analysis related to each research question were interpreted. The results of the analysis for the first research question revealed that there was no statistically significant difference between the learners' scores on computer-based writing and their scores on paper-based testing of writing, which suggests that the medium of test delivery did not bring in difference in the writing scores of the learners across the two modes. The analysis pertinent to the second research question indicated no statistically significant difference between high-computer-familiar group and low-computer-familiar group on computer-based writing. The results provide evidence to conclude that computer familiarity did not introduce construct-irrelevant difference into the writing scores. Finally, based on the evidence provided by the data analysis, the researchers concluded that the two modes comparably measure the same construct

## The Authors

**Mohammad Mohammadi** completed his Ph.D. in TESOL (applied linguistics) in 2004 at the University of Leeds, UK. Since his return to Iran in September 2004, he has been lecturing and researching as an assistant professor at Urmia University. He was also the director of the Scientific and International Relations Office of Urmia University from Nov. 2004 – Jan. 2010. Dr.

Mohammadi's research interests include vocabulary knowledge and vocabulary use, language testing, L2 reading and writing processes, phonetics and phonology, word association techniques, learner autonomy, lexical inferencing, vocabulary learning strategies, listening, pronunciation, writing, speaking and reading strategies, classroom discourse and task-based approaches to language learning and teaching, involving both experimental/quantitative and descriptive/qualitative research.

**Masoud Barzgaran** has got a BA in English language literature from the state university of Urmia. He was admitted to do his MA in TEFL at the same university in 2007. Since 2011, after finishing his MA studies, he has been teaching as an adjunct instructor in Payam Noor University of Urmia.

## References

Chapelle, C. A., and Douglas, D. (2006).Assessing language through computer technology.Cambridge: Cambridge University Press.

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

Fulcher, G. (1999).Computerizing an English language placement test.ELT Journal, 53, 289–99.

Goldberg, A., Russell, M., & Cook, A. (2003). The effect of computers on student writing: A metaanalysis of studies from 1992 to 2002. *Journal of Technology, Learning, and Assessment, 2*(1).

Harrington, S., Shermis, M. D., and Rollins, A. L. (2000).The influence of word processing onEnglish placement test results.*Computers and Composition, 17,* 197–210.

Kim, J. P. (1999). Meta-analysis of equivalence of computerized and P&P tests on ability measures.Paper presented at the Annual Meeting of the Mid-Western Educational Research Association. Chicago, IL.

Lee, H. K. (2004). A comparative study of ESL writers' performance in a paper-based and a computer-delivered writing test. *Assessing Writing*, *9* (1), 4–26.

Lee, Y. -J., (2002). A comparison of composing processes and written products in timed- essay tests across paper-and-pencil and computer modes. *Assessing Writing*, *8*, 135– 257.

Lottridge, S., Nicewander, A., Schulz, M., and Mitzel, H. (2008). Comparability of Paper- based and Computer-based Tests: A Review of the Methodology.

Puhan, P., Boughton, K., & Kim, S. (2007). Examining differences in examinee performance in paper and pencil and computerized testing.*Journal of Technology, Learning, and Assessment, 6*(3).Retrieved April 24, 2010, from http://www.jtla.org.

Taylor, C., Kirsch, I., Eignor, D., & Jamieson, J. (1998).Examining the relationship between computer familiarity andperformance on computer-based language tasks. Language Learning, *49*(2), 219–274.

Wang, S., Jiao, H., Young, M. J., Brooks, T.E., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 assessment: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*, 5-24.

Wang, H., Shin, C. D. (2009). Computer-based & paper-pencil test comparability studies. *Test, measurement, and research services*.Pearson Education.*9*.

# قابلیت قیاس آزمون نگارش کامپیوتری و ورقه‌ای تست PET در بین زبان‌آموزان ایرانی

محمد محمدی
مسعود برزگران
دانشگاه ارومیه

تکنولوژی کامپیوتر به طراحان آزمون این امکان را داده است که بتوانند نسخه کامپیوتری آزمونهای سنتی ورقه ای را بوجود آورند. نسخه های جدید تستهای TOEFL and Cambridge IELTS, BULATS, KET, PET نمونه های خوبی از آزمونهای کامپیوتری هستند. از آنجا که این شیوه نوین آزمون می تواند فاکتورهای تعیین کننده جدیدی از قبیل روش (رسانه) ارائه آزمون، آشنایی با کامپیوتر و غیره را وارد مقوله سنجش زبان کند، این سوال ممکن است پیش بیاید که آیا دو روش آزمون کامپیوتری و ورقه ای بطور قابل مقایسه ای یک توانایی را می سنجند، و آیا می توان نمرات بدست آمده از این دو روش را بعنوان معادل یکدیگر بکار برد؟ در همین راستا تحقیق حاضر قابلیت قیاس نسخه های کامپیوتری و ورقه ای یک آزمون نگارش را بررسی می کند تا مشخص کند که آیا روش (رسانه) ارائه آزمون و نیز میزان آشنایی با کامپیوتر می تواند تغییری نا مرتبط با توانایی مد نظر ( توانایی نگارش) را در نمرات ایجاد کند. داده های این تحقیق از برگزاری بخش نگارش آزمون PET در دو روش کامپیوتری و ورقه ای با 80 نفر زبان آموز ایرانی سطح متوسط بدست آمد. بعلاوه، با استفاده از یک پرسشنامه معتبر آشنایی با کامپیوتر، زبان آموزان به دو گروه "بیش آشنا" و " کم آشنا" با کامپیوتر تقسیم شدند. نتایج آزمون t گروههای مستقل نشان داد که از لحاظ آماری تفاوت معناداری بین نمرات زبان آموزان در آزمون نگارش کامپیوتری و ورقه ای وجود ندارد. نتایج آزمون t گروههای وابسته نیز حاکی از این بود که در در آزمون نگارش کامپیوتری از لحاظ آماری تفاوت معناداری بین نمرات زبان آموزان "بیش آشنا" و " کم آشنا" با کامپیوتر وجود ندارد. بر اساس نتایج تجزیه و تحلیل داده ها و نیز بررسی میزان همبستگی نمرات در دو روش کامپیوتری و ورقه ای اینگونه نتیجه گیری شد که این دو روش (رسانه) ارائه آزمون بطور قابل مقایسه ای یک توانایی را می سنجند.

کلید واژه ها: بررسی قابلیت قیاس، آشنایی با کامپیوتر، توانایی نگارش، آزمون زبان کامپیوتری و ورقه ای