ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 1, No. 6, pp. 935-941, November 2010

doi:10.4304/jltr.1.6.935-941

Comparison between CBT and PBT: Assessment of Gap-filling and Multiple-choice Cloze in Reading Comprehension

Mo Li
School of Foreign Languages, Tianjin University of Technology, Tianjin 300191, China
Email: inklee@126.com

Haifeng Pu
School of Foreign Languages, Tianjin University of Technology, Tianjin 300191, China
Email: inklee@126.com

***Abstract*—The main purpose of the article is to determine whether there is equivalence between computer based test and paper based test. A lot of comparable research has been conducted to investigate the equivalence between two test formats, but their results are inconsistent, which causes controversy in the field of research. Moreover, many of these studies are conducted in the other countries, especially in the USA, but few in China. Therefore, it is necessary to conduct the respective research in the Chinese context. Based on Honaker's two standards for equivalence of computer based test and paper based test, the experiment is made to explore the equivalence between CBT and PBT.**

***Index Terms*—CBT, PBT, psychometric equivalence, experiential equivalence**

I. Introduction

For three decades, educational theorists have proposed many ways in which computers might influence education. Although it was not until the 1970's that computers began having a presence in schools, since then the use of computers in education has increased dramatically. The National Center for Education Statistics reports that the percentage of students in grades 1 to 8 using computers in school more than doubled from 31.5 in 1984 to 68.9 in 1993 (Snyder &Hoffman,1990;1994). Similarly, the availability of computers to students in school increased from one computer for every 125 students in 1983 to one computer for every 9 students in 1995 (Glennan & Melmed, 1996). Now in China, computers are being used more in schools than ever before. In most middle schools, students have easier access to computers.

As the number of computers has increased, theories about how computers might benefit students' reading have proliferated. Actually, some researchers have carried out formal studies to examine whether reading on computer is better than reading on paper. Educational Testing Services (ETS) is already offering the GRE as computer based tests in 180 countries. In 1998, the Educational Testing Services launched CBT TOEFL in the US and numerous countries around the world. In the New Year, the CET committee has decided to adopt the Computer Based Language Testing (CBLT) on Band4.6 tests. So just as Educational Testing Services (1997:2) predicted "computerized assessment might have been the 'road less traveled' in the early 90s, but today it is the future of testing."

II. Background

Two major computer testing procedures have been introduced: computerized adaptive testing (CAT) and computer based testing (CBT). To facilitate the present study, these two testing procedures will be differentiated. Computer adaptive testing refers to a sophisticated form of testing, where the computer dynamically selects items to administer to a given examinee based upon his/her earlier item responses on the test. Most proposed systems for developing adaptive tests are based on the item response theory, and most adaptive tests are primarily used in ability and achievement tests (Ju, 1993). Computer based testing as applied in this study, however, generally refers to using a computer to give exactly the same test as one in a paper and pencil format. That is, it has the same test questions and presents them in exactly the same order as the paper and pencil version of the test. Most computer based testing was developed based on the classical test theory, and have been extensively studied (Dunn, Lushene, &O'Neil, 1972; Elwood, 1972; Kiely, Zara, &Weiss, 1986). Computer based testing provides many advantages, such as greater standardization, reduced testing time, immediate test results, and the ability to measure item response latencies. With the increasing availability and utilization of microcomputers in educational institutions, the market for computer based tests will continue to grow.

As computer based tests become available, the equivalence between computer-generated scores and corresponding paper-pencil scores becomes a critical issue. Criteria for the equivalence between two formats of a test include

psychometric equivalence and experiential equivalence (Honaker, 1988). The psychometric equivalence depends on the format of administration effect involving mean scores, distributions and correlations with scores between two test formats under the consideration of methodologies. If computer-based test and paper-pencil test results come up with these standards, then validity data from one format can be directly generalized to the other. Psychometric equivalence is considered to be the primary concern in determining the equivalence between two test formats. Moreover, the other concern involves the evaluation of experiential equivalence, based on the influence of individual difference on the computer-based test involving the factors such as computer experience and computer anxiety. If these factors distract the result of computer-based test, then the equivalence between two test formats are threatened.

Many testing programs are increasingly administering the same tests in both PBT and CBT formats. For example, the TOEFL program concurrently delivers PBT and CBT in approximately 228 countries every year. Similarly, the GRE General test is offered in the US, Canada, and many other countries in paper-pencil and computer-based formats. Mills, Potenza, Fremer, and Ward (2002) speculate that this trend will continue to increase, because of an increase in availability of microcomputers in educational settings, a substantial improvement in the speed of computers, and a significant reduction in cost. CBT has many advantages over PBT, which include faster score reporting, savings on paper and personnel resources and costs of scoring services (Wise & Plake, 1990), and development of new methods of assessment such as simple adaptations of multiple-choice items to more innovative item types (Jodoin, 2003). Despite these advantages, an important question that arises when tests are administered in both formats is whether or not the scores produced are interchangeable (Wang & Kolen, 2001; Gallagher, Bridgeman, & Cahalan, 2002). For example, scores derived from CBT as compared to PBT might reflect not only the examinee's proficiency on the construct being measured, but also differences in formatting (including typing vs. hand-writing) and/or computer proficiency.

There is a large body of research that documents the comparability of scores obtained from PBT and CBT. Lee (1986) investigated the relationship of past computer experience on the scores of college students on a computer-based arithmetic reasoning test and concluded that students with little or no computer experience scored significantly lower than students who had previous experience with computers. Similarly, an extensive review by Mazzeo and Harvey (1988) found that CBT tended to be more difficult than PBT versions of the same tests. Furthermore, Mead and Drasgow (1993) found that the constructs being measured across the two modes were similar for power tests but not for speeded tests. However, other researchers have found that PBT and their CBT counterparts yield comparable scores. For example, Taylor, Jamieson, Eignor, and Kirsch (1998) studied the comparability of PBT and CBT for the 1996 administration of the TOEFL and found no meaningful difference in performance for examinees taking the two different versions. Similarly, Wise, Barnes, Harvey, and Plake (1989) contend that PBT and CBT versions of achievement tests yield very similar scores.

Since the results of these studies are inconsistent and the use of computers have become commonplace, it is even more important to examine whether scores obtained from the two different modes of delivery are in fact comparable (Gallagher, Bridgeman, & Cahalan, 2002). In addition, most of the research studies on the comparability of CBT and PBT were conducted during the 80s and 90s, but the computer technology develops on the monthly and even daily basis in recent years, the conditions for the CBT are changed greatly. Moreover, many research studies related to CBT are conducted in other countries, but in most of the educational institutions in China, especially in primary schools and middle schools, the influence of the computer's application in language tests have not been seen. Therefore, it is significant that some comparable studies related to CBT and PBT be conducted to verify those research reports in our country. So many factors should be reconsidered and new ones should be included in the following-up researches. Under the changing conditions, the current research analyzes the correlation of the two versions of tests and the factors involved.

III. Methodology

A. Subjects

The subjects are Grade Two middle school students from one class, Pingdingshan Feixing School. The whole class, which contains 46 students, is split half randomly to form two groups. They are labeled Group A and Group B. Group A consists of 23 students. Group B consists of 23 students. The age range of the students at the time of the test is from 13 to 16 years old with a mean of 14.5 years old. From the two groups, valid data samples are collected from CBT and PBT respectively. To achieve the research goals, this study will employ a quantitative analysis approach including descriptive statistics, correlation analysis etc.

B. Test Measures

A single passage is used for the two versions of reading comprehension measurement developed in this study. The text "The Cock Crows at Midnight", whose Chinese version the candidates are familiar with, has ever been taught in the Chinese textbook in the primary school. The text is very humorous and interesting, as long as the subjects read the beginning of the text, most of them can guess the meaning of the whole test. This enhances the readability of the text, which can activate the reader's interest. It can avoid the case that the candidates often over-focus on the mechanics of the cloze and neglect the content-- a forest and trees issue in the past reading comprehension.

The text is designed to delete the words randomly, which is easy-understanding for the middle school students. The

resulted passages are formed into two types of tests, i.e. the multiple choice test and the gap-filling test with respective eleven blanks. The former contains eleven blanks without any distracters and the latter eleven blanks with three distracters. The two types of tests are also fashioned into their respective computer and paper formats.

A questionnaire designed to measure computer anxiety and experience was developed for this study and administered to each of the subjects of the gap-filling and multiple-choice cloze tests. The questionnaire consists of 10 questions concerning attitudes towards computers, and 4 questions concerning previous computer experience. The attitudes questions are answered on a 5-point scale determined by degree of agreement, the lower the score, and the greater the degree of computer anxiety. The experience questions are rated on a 4-point scale, the higher the score, and the greater the degree of experience (See appendix8 for the attitude and experience scale)

C. Test Procedure

The experiments are made between the two groups, which are chosen from the same class in the middle school. First, Group A and Group B take the computer based test respectively. Group A takes the computer based gap filling cloze test first, while Group B takes the computer based multiple-choice cloze test. No time limited is set during each test, but time consuming on the test will be recorded on the computer when the candidates finish the test. They only have to click the 'Check' button to upload their test results.

A week later, the PBT will be taken by the subjects. Group A takes the PBT gap filling cloze test, Group B takes the multiple-choice cloze test. No time is set. After the candidates finish their paper tests, their time consuming on the tested will be recorded by the test administers.

D. Research Hypothesis

1. There is a significant difference between scores on paper based gap-filling cloze test and computer based gap filling cloze test.
2. There is a significant difference between scores on paper based multiple-choice cloze test and computer based multiple-choice cloze test.
3. The computer anxiety affects computer based test scores.
4. The computer experience affects computer based test scores.
5. Age affects students' computer based test scores.
6. Gender affects students' computer based test scores.

IV. DISCUSSION

As Gallagher et al. (2002) pointed out, with an increase in familiarity of students with computers, an overall measure of difference in test performance due to change in mode of delivery may appear less meaningful today. Thus, it is important to use both statistical and substantive analyses at the test and item level in order to ensure that tests are fair and valid for all, regardless of mode of presentation. So the statistical package SPSS 11.5 for Windows is used to analyze the data in this experiment. As evident, the findings of this study are positive and suggested that the CBT and PBT versions of the reading comprehension are comparable.

A. Descriptive Statistics

Means and standard deviations for the variables are shown in Table 4.1. The mean test scores for the paper-based gap-filling cloze test and computer-based gap-filling cloze test are 17.48 and 16.74, with standard deviations of 2.84 and 3.23 respectively, indicating similarly of performance for the two versions of the gap filling cloze test. The mean test scores for the paper based multiple-choice cloze test and the computer based multiple-choice cloze tests are 18 and 17.3, with standard deviations of 3.46 and 3.78 respectively. While these also indicate a similarity between the two versions of multiple-choice cloze test, the mean performance for the multiple-choice cloze test is a little higher than that of the gap filling cloze test, and the dispersion of scores for the multiple-choice cloze test is higher than that of the gap filling cloze test.

TABLE 1
MEAN SCORE AND STANDARD DEVIATION

Gap-filling test	N	Min	Max	Mean	Std. Deviation
PBT	23	12.00	22.00	17.4783	2.84237
CBT	23	8.00	18.00	16.7434	3.23650

TABLE 2
MEAN SCORE AND STANDARD DEVIATION

Multiple-choice test	N	Min	Max	Mean	Std. Deviation
PBT	23	12.00	22.00	18.0000	3.46410
CBT	23	8.00	22.00	17.2609	3.78045

The mean score for the attitude scale is 32.4 (maximum=50), indicating a generally positive attitude towards computers. A score of 30 or above might be regarded as signifying a positive attitude towards computers, while a score

below 20 signifies a negative attitude or computer anxiety. The questionnaire (See Appendix 8) shows that none of the subjects can be described as computer anxious, and only one as equivocal (subject with scores of 23), with the rest having positive attitudes.

The mean score for computer experience is 9.9 (maximum score=16), suggesting a reasonable degree of computer experience. A subject with a score below 6 might be considered to have little computer experience, from 7 to 12 to have some experience, and from 13 to 16 to be very experienced. Questionnaire B shows that only three subjects with score of 5.5 and 6 have little experience about computer, most of the students have much experiences, and they have higher score for question 1,2 and 3; question 4 scores were much lower. This indicates the subjects' general lack of experience on the Internet.

B. Correlation Analysis

TABLE 3
2-TAILED T-TEST OF THE CORRELATIONS OF 4 VARIABLES

		computer attitude	computer experience	cbt multiple-choice cloze test score
computer attitude	Pearson Correlation	1	-.065	.244(**)
	Sig. (2-tailed)	.	.207	.000
	N	374	374	374
computer experience	Pearson Correlation	-.065	1	.060
	Sig. (2-tailed)	.207	.	.249
	N	374	374	374
cbt multiple-choice cloze test score	Pearson Correlation	.244(**)	.060	1
	Sig. (2-tailed)	.000	.249	.
	N	374	374	374

Table 3 gives the correlations between the computer-based cloze test and subjects' attitude and experience. Correlations among computer based multiple-choice cloze test, computer attitude and computer experience are .244 and .06 respectively. P values are .00 and .244. This shows that there is no significant correlation between the computer attitude and computer-based multiple-choice cloze test score ($r=.244$, $p<.01$), and there is also no correlation between the computer experience and computer-based multiple-choice cloze test score ($r=.06$, $p>.01$)

TABLE 4.
2-TAILED T-TEST OF THE CORRELATIONS OF 4 VARIABLES

		computer attitude	computer experience	cbt multiple-choice cloze test score
computer attitude	Pearson Correlation	1	-.065	.244(**)
	Sig. (2-tailed)	.	.207	.000
	N	374	374	374
computer experience	Pearson Correlation	-.065	1	.060
	Sig. (2-tailed)	.207	.	.249
	N	374	374	374
cbt multiple-choice cloze test score	Pearson Correlation	.244(**)	.060	1
	Sig. (2-tailed)	.000	.249	.
	N	374	374	374

Correlations between the computer-based gap filling cloze test and computer attitude, computer experience are shown in Table4.2b. Pearson's product correlation are respectively.182 and .151, P values are .001 and .007. This shows that there is no significant correlation between the computer attitude, computer experience and computer-based gap filling cloze test.

TABLE 5
2-TAILED T-TEST OF THE CORRELATIONS OF 4 VARIABLES

Correlations

		computer experience	computer attitude	cbt gap-filling cloze test score
computer experience	Pearson Correlation	1	-.196**	-.182**
	Sig. (2-tailed)	.	.000	.001
	N	316	316	316
computer attitude	Pearson Correlation	-.196**	1	.151**
	Sig. (2-tailed)	.000	.	.007
	N	316	316	316
cbt gap-filling cloze test score	Pearson Correlation	-.182**	.151**	1
	Sig. (2-tailed)	.001	.007	.
	N	316	316	316

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

TABLE 6
2-TAILED T-TEST OF THE CORRELATIONS OF 5 VARIABLES

Correlations

		age	gender	CBT multiple-choice score	Computer Attitude	Computer Experience
AGE	pearson correlation	1	-.168	-.382	.025	-.122
	Sig. (2-tailed)	.	.432	.072	.909	.571
	N	24	24	23	24	24
GENDER	pearson correlation	-.168	1	.008	.123	-.656**
	Sig. (2-tailed)	.432	.	.970	.568	.001
	N	24	24	23	24	24
	pearson correlation		**			

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

TABLE 7
2-TAILED T-TEST OF THE CORRELATIONS OF 5 VARIABLES

Correlations

		age	gender	computer attitude	computer experience	CBT gapfilling scores
age	pearson correlation	1	-.070	-.115	.035	-.285
	Sig. (2-tailed)	.	.756	.609	.877	.199
	N	22	22	22	22	22
gender	pearson correlation	-.070	1	.255	-.546**	-.414
	Sig. (2-tailed)	.756	.	.253	.009	.056
	N	22	22	22	22	22
	pearson correlation		**			

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

Table 6 and 7 shows that there is no significant correlation between the age and computer attitude; computer experience; and computer based formats scores. Likewise, there is no significant correlations between gender and computer attitude; computer-based formats, however, there is a significant middle correlation between computer experience and gender (Pearson correlation=0.6, p=.009)

Correlations between the paper-based gap filling cloze test and computer-based gap filling cloze test is .71. There is a significant correlation between the paper-based gap filling cloze test and computer-based gap filling cloze test (p= .0093).

Correlation between the paper-based multiple-choice cloze test and computer-based multiple-choice cloze test is .75. There is a significant correlation between the paper-based multiple-choice cloze test and computer-based

multiple-choice cloze test (p=.00). These correlations are shown in Table 4.4

TABLE 8
CORRELATIONS EFFICIENTS BETWEEN CBT AND PBT

	Test score	correlation	Sig.
Group A	CBT gap-filling cloze test & PBT gap-filling cloze test	.712	.0093
Group B	CBT multiple-choice cloze test &PBT multiple-choice cloze test	.750	.00

V. CONCLUSION

The main purpose of this study is to determine whether the computer based gap-filling cloze and multiple-choice cloze tests are suitable tools for measuring reading comprehension, by comparing students' performance on traditional style paper-based tests to performance on the same tests of computer-based format. From a psycholinguistic viewpoint, the language communication model suggests that the ability of a reader to decode text is affected by the message system. Thus, when a computer screen replaces the traditional paper-based presentation of a text, the comparative differences in the message system may change the readers' decoding or comprehension of that text. Previous researchers have noted that when the subjects do computer based testing, their performance may be influenced by factors intrinsic to the computer mode, such as screen clarity and ease of navigating the text, or by personal factors such as anxiety or inexperience with computers.

While much research has been done in the area of computer-based testing, there are also some studies on the suitability of computer-based testing. In this study , four kinds of reading comprehension measurements are designed and administered to Grade Two 46 subjects, along with measures of attitude towards computers and computer experience, to explore whether the computer based testing is a suitable testing medium and whether factors such as computer anxiety or experience can affect test performance. Statistical analysis of the test scores provides answers to the research questions presented in Chapter4.

1. There is no significant difference between the paper based gap filling cloze test and computer based gap filling cloze test.
2. There is no significant difference between the paper based multiple-choice cloze test and computer based multiple-choice cloze test.
3. Computer anxiety does not affect students' computer based test scores.
4. Computer experiences do not affect students' computer based test scores
5. Age does not affect students' computer based test scores.
6. Gender does not affect students' computer based test scores.

The study indicates that the students achieve similar test scores for the computer based gap filling cloze test and paper based gap filling cloze test, and similar for the computer based multiple-choice cloze test and paper based multiple-choice cloze test, and also the dispersions of the two versions of the gap filling and multiple-choice cloze test are similar, so it can prove that computer based gap filling cloze test is equivalent to the paper based gap filling cloze test, and the computer based multiple-choice cloze test is equivalent to the paper based multiple-choice cloze test according to American Psychological Association (APA) guideline.

Furthermore, no significant relationship is found between computer anxiety and computer based test scores, and also no significant relationship is found between computer experience and computer based test scores. In other words, the student in this study who has higher computer anxiety does not score lower or spend more time on the computer based testing than subjects who have lower computer anxiety. Students who have more computer experience do not score higher, and spent less time on the computer based testing than students who have less computer experience. In this study, because students generally have low computer anxiety even though they have little experience, computer anxiety does not seem to influence test performance in computer based testing. Even if most of the subjects have much experience, it does not seem to influence their performance in the computer based testing. However, in this study, it is found that the students who have little computer experience may tend to spend more time completing a computer based test than who have much computer experience.

In general, the subjects in this study show positive attitudes toward computer based testing; although they have never experienced or heard about computer based testing until participating in this experiment. It is believed that positive attitudes toward computer based testing will help the development of computer based testing and education from screen reading. Based on the findings of this research, the investigated variables-computer anxiety or computer experience does not result in the two different format administrations. But there are still engineering designs or technical problems that maybe result in the differences between computer-based testing and paper based testing, such as the program, the computer itself or interface design, also could be the factors to make the two test formats different.

REFERENCES

[1] Dunn, T.G., Lushene, R.E., & O'Neil, H.F. (1972). Complete automation of the MMPI and a study of its response in latencies.

Journal of Consulting and Clinical Psychology.39,381-387.
[2] Elwood, D.L. (1972). Automated WAIS testing correlated with face-to-face WAIS testing: A validity study. *International Journal of Man-Machine Studies,4.*129-137.
[3] Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The effect of computer-based tests on racial-ethnic and gender groups. *Journal of Educational Measurement*, 39 (2), 133-147.
[4] Glennan, T. K., & Melmed, A. (1996). Fostering the use of educational technology: Elements of a national strategy. Santa Monica, CA: RAND
[5] Honaker, L.M. (1988). The equivalency of computerized and conventional MMPI administration. *Clinical Psychology Review*, 8: 561-577.
[6] Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, 40 (1), 1-15.
[7] Ju, Gin-Fon Nancy. (1993). A computer-based Chinese edition of DAT mechanical reasoning: Comparing computer-based and paper-pencil formats of a timed pictorial tests in Taiwan. University of Illinois at Urbana-Champaign.
[8] Kiely, G. L., Zara, A. R., & Weiss, D. J. (1986). Equivalency of computer and paper-and-pencil Armed Services Vocational Aptitude Battery tests. Brooks Air Force Base, TX: Air Force Human Resources Laboratory.
[9] Lee, J. (1986). The effects of mode of test administration on test performance. *Educational and Psychological Measurement.*46:467-474.
[10] Mazzeo, J & Harvey, A.L. (1988). The equivalence of scores from conventional and automated educational and psychological tests: A review of literature. Princeton. NJ: Educational Testing Service
[11] Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449-458.
[12] Mills, C. N., Potenza, M.T., Fremer, J.J., & Ward, C. W. (2002). Computer-based testing: Building the foundation for future assessments. Mahwah, NJ: Lawrence Erlbaum.
[13] Taylor, C., Jamieson, J., Eignor, D. R., & Kirsch, I. (1998). The relationship between computer familiarity and performance on computer-based TOEFL test tasks (ETS Research Report No: 98-08). Princeton, NJ: Educational Testing Service.
[14] Snyder, T. D., & Hoffman, C. M. (1994). Digest of Education Statistics, 1994. Washington, DC: National Center for Education Statistics, U.S. Department of Education. ED 377 253.
[15] Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement,* 38 (1), 19-49.
[16] Wise, S.L., Barnes, L. B., Harvey, A. L., & Plake, B. S. (1989). Effects of computer anxiety and computer experience on the computer-based achievement test performance of college students. *Applied Measurement in Education*, 235-241.
[17] Wise, S. L., & Plake, B. S. (1990). Computer-based testing in higher education. *Measurement and evaluation in counseling and development*, 23 (1), 3-10.

Mo Li was born in Pingdingshan, Henan Province in 1976. She received her M.A. degree in linguistics from Henan Normal University, China in 2005. She is currently a lecturer in School of Foreign Languages, Tianjin University of Technology. Her research interests include teaching methodology and English language testing.

Haifeng Pu was born in Heilongjiang Province in 1975. He received his M.A. degree in translation from Henan Normal University, China in 2005. He is currently a lecturer in School of Foreign Languages, Tianjin University of Technology. His research interests include translation and English literature.