English Language Teaching; Vol. 13, No. 1; 2020
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

A Comparative Study of Test Takers' Performance on Computer-Based Test and Paper-Based Test Across Different CEFR Levels

Don Yao[1]

[1] The Language Assessment Seminar Research Group, Department of English, Faculty of Arts and Humanities, University of Macau, Macau, China

Correspondence: Don Yao, The language assessment seminar research group, Department of English, Faculty of Arts and Humanities, University of Macau, Macau, China.

Received: December 4, 2019 Accepted: December 19, 2019 Online Published: December 20, 2019

doi: 10.5539/elt.v13n1p124 URL: https://doi.org/10.5539/elt.v13n1p124

Abstract

Computer-based test (CBT) and paper-based test (PBT) are two test modes to the test takers that have been widely adopted in the field of language testing or assessment over the last few decades. Due to the rapid development of science and technology, it is a trend for universities and educational institutions striving rather hard to deliver the test on a computer. Therefore, research on the comparison between these two test modes has attracted much attention to investigate whether the PBT could be completely replaced. At the same time, task difficulty is always a key element to reflect test takers' performances. Numerous studies have laid a solid foundation and guidance about the comparative study of test takers' performance on CBT and PBT, but there still remains a scarcity from the perspective of task difficulties with different Common European Framework of Reference for Languages (CEFR) task levels in particular.

This study, therefore, compared the test takers' performance on both CBT and PBT across tasks with different CEFR levels. A total of 289 principal recommended high school test takers from Macau took the pilot Test of Academic English (TAE) at a local university. The results indicated that there was a difference between test takers' performance on different test modes across different CEFR levels, but only CEFR A2 level showed a statistically difference between CBT and PBT. And since science and technology are continuously developing, it is essential for the university to consider switching the test mode from PBT to CBT.

Key words: comparative, paper-based test, computer-based test, CEFR levels

1. Introduction

In response to the rapid advancement of science and technology, universities and educational institutions are striving very hard, with the latest software on hand, for a developed new assessment method to replace their previous test mode. Computer-based test (CBT), a test or an assessment administered by a computer, has incrementally grown since the 1970s. And it is now widely employed by various kinds of tests, especially for large-scale institutional tests like TOEFL, IELTS, or a university placement test, etc.

CBT, compared with paper-based test (PBT), is less time-consuming, easier and quicker to be administered and scored, etc. However, it is also faced with some technical challenges such as the compatibility of software. The past few decades have witnessed an upsurge of interest in the research on comparison between CBT and PBT. Some studies showed that test takers' performance had a significant difference between these two test modes due to gender, age, familiarity with the computer, etc. (Parshall & Kromrey, 1993; Gallagher, Bridgeman & Cahalan, 2000; Oduntan, Ojuawo & Oduntan, 2015; Choi &Tinkler, 2002; Wang et al., 2008; Goldberg & Pedulla, 2002; Jeong, 2008). However, some other studies revealed the opposite result that test takers' performance had no significant difference between these two test modes (Mazzeo & Harvey, 1988; Mead & Drasgow, 1993; Anakwe, 2008; Öz & Özturan, 2018). These studies have laid a solid foundation and guidance about the comparative study of test takers' performance on CBT and PBT. However, task difficulty as a key element in a test can always reflect test takers' performance, but nothing has been done from the perspective of task difficulties with different Common European Framework of Reference for Languages (CEFR) task levels in particular.

Given the above-mentioned facts, this study aims to investigate whether there is any difference between CBT and PBT across tasks with different CEFR levels based on test takers' performance. The results are able to bring about the positive influence on universities or educational institutions to use a more suitable test mode to deliver the test to achieve educational or even pedagogical purpose. As the technology develops, it is important to switch the test from the paper-and-pencil mode to the computer-based mode. And it is essential to offer an innovative online test that would be a model test in a university setting. A total of 289 principal recommended high school test takers from Macau took the pilot Test of Academic English (TAE) at a local university, among which 53 took the CBT and 236 took the PBT. Moodle platform was used for CBT. Their test scores were scored automatically by Moodle. The language assessment for the listening and reading sections and rated by two experienced staff for the writing section. Content and statistical analyses were conducted to understand the linguistic features of the tasks and test performance of the test takers. Coh-Metrix, Flesch Kincaid Grade Level, LexTutor and Statistical Package for the Social Sciences (SPSS, version 24.0) were used for these analyses.

2. Literature Review

The comparison between CBT and PBT has been researched for the last few decades. Some studies were in favor of that test takers' performance had a significant difference between these two test modes in terms of the variables like gender, age, familiarity with the computer, etc.

2.1 Test Modes had Influence on Test Takers' Performance

2.1.1 Gender

Parshall and Kromrey (1993) investigated 1114 test takers who took Graduate Record Examination (GRE) to see whether their performance was closely related to the test modes or not. A T-test was used to examine the relationship between gender and test takers' performance on both PBT and CBT. Overall, test takers performed better on CBT than PBT. In terms of gender, the result showed that male test takers performed better on CBT than PBT.

Gallagher, Bridgeman, and Cahalan (2000) examined several national test programs including GRE, Graduate Management Admissions Test (G-MAT), etc. to see whether the change from PBT to CBT had influence on test takers' performance. They concluded that all the differences were quite small except for the racial-ethic and gender groups. Hispanic test takers did slightly better than Africa-American test takers. And male test takers performed slightly better than female test takers.

Oduntan, Ojuawo and Oduntan (2015) did research on test takers in Nigeria who took the Unified Tertiary Matriculation Examination (UTME) in both 2013 and 2014 by offering them a questionnaire. Pearson correlation was used for descriptive analysis. They found that test takers did better on CBT than the same test takers who did PBT. And female test takers outperformed than male test takers. The positive correlation in the scores of test takers indicated they were gradually interested in CBT. And if they better prepared for the CBT, their performance would be better.

2.1.2 Age

Choi and Tinkler (2002) evaluated the comparison between CBT and PBT to Grade 3 and Grade 10 test takers in a mathematics and reading test to see whether age had influence on test scores. They found that CBT had greater effect on Grade 3 test takers than Grade 10 test takers. They thought it might because the function of scrolling on the computer, especially when test takers did the reading part, was more favorable of Grade 10 test takers.

Wang et al. (2008) conducted a meta-analysis to examine whether test modes had influence on K-12 student reading assessment. K-12 is a term that represents students from kindergarten to Grade 12. The result showed that the variable grade level was statistically significant, and the higher the grade was, the better test takers did on CBT than PBT. Some other variables like study design, sample size, type of test, etc., none were statistically significant.

2.1.3 Familiarity with the Computer

Goldberg and Pedulla (2002) did research on 222 test takers with different computer familiarities (low, moderate and high) who took GRE in Boston. Their idealism was to find there was no difference between two test modes based on test takers' performance so that they could transfer the test mode from PBT to CBT. However, they found that test takers took PBT outperformed than that test takers took CBT on the whole. For the CBT, test takers highly familiar with the computer did better than those who lowly familiar with the computer.

Jeong (2008) investigated 73 test takers in a public school in Korea. They were asked to finish the test by both test modes. And the test was composed of four subjects: Korean language, Mathematics, Social studies and Science. The results showed there were statistically significant difference in just Korean language and Science. And test takers outperformed on PBT than CBT. This result was surprising because according to Jeong, these test takers were very familiar with the computer. However, they did not achieve higher scores on CBT. Thus, the familiarity with the computer as a variable indeed influenced the final result, but it was not always a positive result.

2.1.4 Test Modes had no Influence on Test Takers' Performance

On the other hand, some other studies indicated that test takers' performance had no significant difference between these two test modes. Mazzeo and Harvey (1988) combined several literature reviews, with 30 comparability studies included, laying emphasis on different aspects like personality, aptitude, intelligence, etc. And they found different test modes had no effect on power tests but speeded test.

Mead and Drasgow (1993) used meta-analytic techniques to examine the correlation across formats. A total of 159 correlations were computed with 123 from power tests and 36 speeded tests. The overall correlation between CBT and PBT was .91 out of 1.00. The correlation for power tests was .97 but for speeded tests was only .72 which indicated the similar result that Mazzeo and Harvey got in 1988.

Anakwe (2008) took gender and class as two variables to see whether they would affect test takers' performance on different test modes. A two-way ANOVA was used to take gender and class as independent variables and test takers' scores as dependent scores to determine the final result. He found there was no significant difference between CBT and PBT on test takers' performance.

Öz and Özturan (2018) investigated 97 Turkish test takers who enrolled the English as a foreign language (EFL) teacher program. They took both CBT and PBT. The result indicated that there was no any impact on the reliability and validity of the test modes in either way. And they also found that there was no significant difference between test scores and test modes.

As mentioned before, these studies were significant bases for the comparative study of test takers' performance on CBT and PBT. But researchers may ignore the importance of task difficulties, especially for tasks at different CEFR levels. Therefore, two research questions were articulated in this study:

Research question 1: *Is there a difference between test takers' performance on CBT and PBT?*

Research question 2: *Is there a difference between test takers' performance on CBT and PBT across tasks at different CEFR levels?*

3. Methodology

3.1 Participants

A total of 289 study participants took part in this study, among which 53 took the CBT and 236 took the PBT. The study participants were the students recommended by the principals of Macau high schools called PRA students with the following conditions: (1) Excellent academic performance and good conduct; (2) Currently studying in Form 6 (high school senior year) equivalent in an officially recognized secondary school in Macau; (3) Ranking among the top 10% in terms of academic performance amongst all Form 6 students or equivalent in the school recommended by Macau secondary school principals. And the results of this test would influence their placement in the English levels.

3.2 Instruments

Two instruments were used in this study: (1) CBT of TAE, which was a test constructed on Moodle platform with additional plug-ins; (2) PBT of TAE, which was adopted from the online version of TAE. Both versions had three sections: Section 1 (reading), Section 2 (listening), and Section 3 (writing). The reading section had seven parts (40 items) with selected responses, the listening section had four parts (26 items) with both selected and constructed response formats, the writing section had two parts, an independent e-mail writing and a reading-writing integrated part that included pre-writing questions and essay writing.

Also, all the tasks were at different CEFR levels from CEFR A2 level to CEFR B2 level. For reading section, Part 1, 2 and 3 were at CEFR A2 level, Part 4 and 5 were at CEFR B1 level, and Part 6 and 7 were at B2 level; for listening section, Part 1 was at CEFR A2 level, Part 2 and 3 were at CEFR B1 level, and Part 4 was at CEFR B2 level; for writing section, the E-mail writing was at CEFR A2 level, the three pre-academic questions were at CEFR B1 level, and the essay writing was at B2 level.

3.3 Administration

CBT of TAE was administered in four computer labs at ELC. The labs were equipped with desktop computers and speakers. The computers had a safe browser installed for security measures. Test takers were provided with special Moodle accounts to log in. Before the test, test takers were provided with a tutorial video (5 minutes) on how to take the test (instructions and functions) that they could watch as many times as they want. After the reading section, the listening section was group administered through the speakers, followed by writing. The reading section was 40 minutes, the listening section was 20 minutes and the writing section was 55 minutes. The reading and writing sections were set up to end the test after the limited time period and collect the answers automatically. Test takers could not go back to the completed sections after the set-up time period.

PBT of TAE was administered in eight classrooms equipped with the same speakers as in the computer labs. Test takers were given out the test booklets, answer sheets for reading and listening, and answer sheets for writing. After the reading section, the listening section was group administered through the speakers, followed by writing. The sections took the same amount of time to complete, and test takers were not allowed to go back to the completed sections after the set-up time period.

3.4 Data Collection

The test performance data of the reading and listening sections was scored dichotomously (0 for incorrect response and 1 for correct response for all items). The scoring of the CBT was done automatically by Moodle. The results were downloaded as an Excel file, which was feasible for data analysis. The exception was the *Fill-out the Form* task. Due to the task type, the items were embedded into two tables, one included seven items and the other included four items. Each of the table was considered as a single item by Moodle. Therefore, data retrieved from Moodle only showed the total scores of the whole part. To get the scores for each item separately, the test items were reviewed and the data were input manually.

For the PBT, the answer sheets were scanned by the Scantron. The data of the multiple choice was automatically downloaded into an Excel file. The *Fill-out the Form* task was scored manually by the research assistants (RAs) at the university.

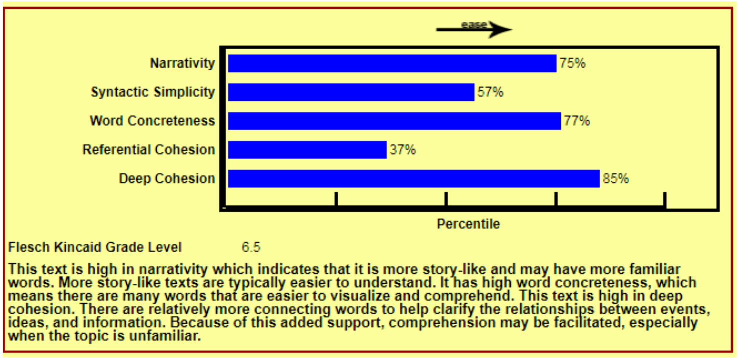
Both CBT and PBT of the writing section were manually scored by instructors from ELC and TAE project RAs based on in-house rating rubrics. The rating rubrics drew on the construct definition of the writing tasks and/or existing rubrics of international and local English language tests (e.g., IELTS, TOEFL, and JAE). Holistic scoring was used for rating e-mail and essay, and analytical scoring was used for rating pre-writing questions.

3.5 Data Analysis

3.5.1 Coh-Metrix and Flesch Kincaid Grade Level

The computational tool, Coh-Metrix, was used to examine whether the difficulty of texts was suitable for test takers at those different CEFR levels in terms of lexical, syntactic and discourse features. Some indices like narrativity, syntactic simplicity, word concreteness, referential cohesion and deep cohesion were got to help with the final results. Flesh Kincaid Grade Level revealed the Grade level fits the U.S. students. According to the British Council, CEFR A2 level is equal to Grade 6 to 7 students, CEFR B1 level is equal to Grade 8 to 9 students, and CEFR B2 level is equal to Grade 10 to 12 students.

Figure 1, 2 and 3 display the score and level for CEFR A2, B1 and B2 level texts. Figure 1 presents the score and level for CEFR A2 level text. The text is high in narrativity which means it is very story-like and may have many familiar words. It also has high in word concreteness, which indicates most of the words are easy to visualize and comprehend. The text is also high in deep cohesion, which means there are relatively more connecting words to help combine the relationships between events or ideas. And the Flesch Kincaid Grade level is 6.5 which means the difficulty is suitable for test takers at CEFR A2 level.

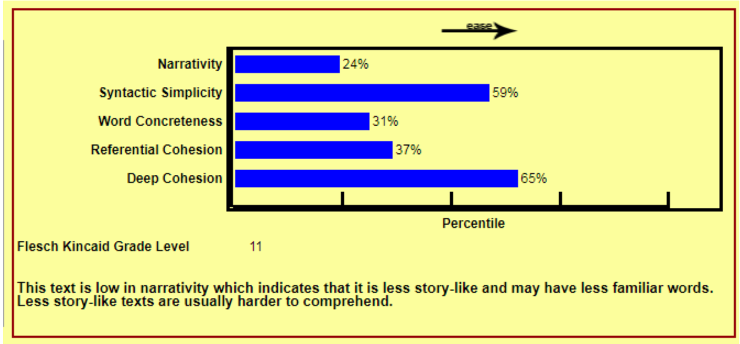


Flesch Kincaid Grade Level 6.5

This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text is high in deep cohesion. There are relatively more connecting words to help clarify the relationships between events, ideas, and information. Because of this added support, comprehension may be facilitated, especially when the topic is unfamiliar.

Figure 1. Score and level for CEFR A2 level text

Figure 2 shows the score and level for CEFR B1 level text. The text is high in syntactic simplicity, which means it has simple sentence structures. It is really low in word concreteness, which indicates the text is full of abstract words or phrases. And the Flesch Kincaid Grade level is 8.8 which means the difficulty is suitable for test takers at CEFR B1 level.

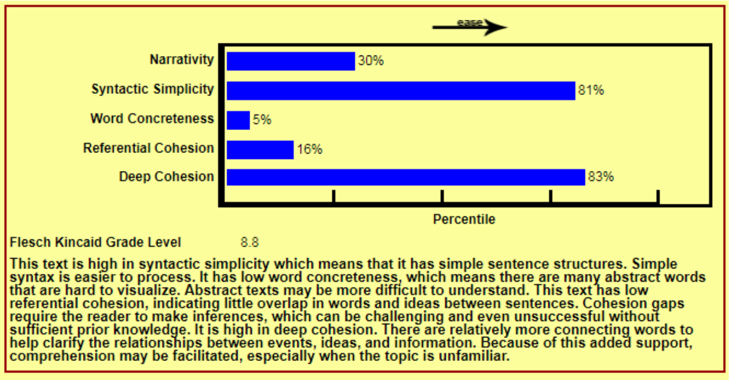


Flesch Kincaid Grade Level 11

This text is low in narrativity which indicates that it is less story-like and may have less familiar words. Less story-like texts are usually harder to comprehend.

Figure 2. Score and level for CEFR B1 level text

Figure 3 displays the score and level for CEFR B2 level text. The text is low in narrativity, which means it is not much story-like and may have very few familiar words. It is also low in referential cohesion, which indicates it has very few overlaps in words or ideas between sentences. And the Flesch Kincaid Grade level is 11.0 which means the difficulty is suitable for test takers at CEFR B2 level.



Flesch Kincaid Grade Level 8.8

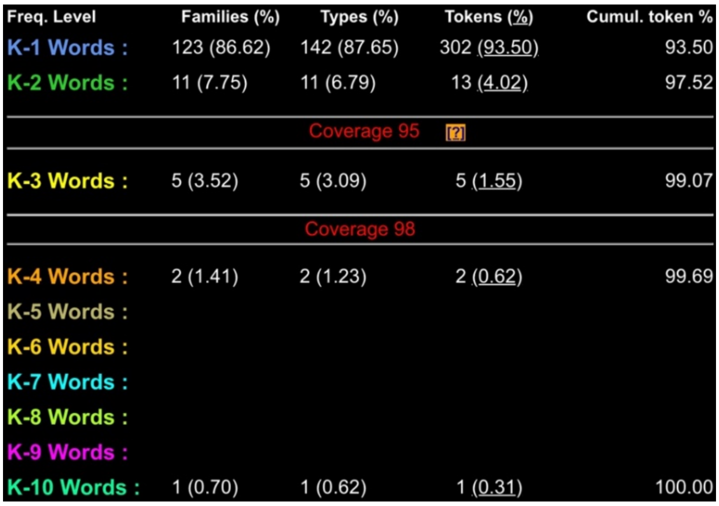
This text is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. It has low word concreteness, which means there are many abstract words that are hard to visualize. Abstract texts may be more difficult to understand. This text has low referential cohesion, indicating little overlap in words and ideas between sentences. Cohesion gaps require the reader to make inferences, which can be challenging and even unsuccessful without sufficient prior knowledge. It is high in deep cohesion. There are relatively more connecting words to help clarify the relationships between events, ideas, and information. Because of this added support, comprehension may be facilitated, especially when the topic is unfamiliar.

Figure 3. Score and level for CEFR B2 level text

3.5.2 LexTutor

LexTutor was used to examine whether the difficulty of lexical words was suitable for test takers at those different CEFR levels. Also, according to the British Council, for test takers at CEFR A2 level, they should master K-3 words; for CEFR B1 level, they should master K-4 words; and for CEFR B2 level, they should master K-5 words.

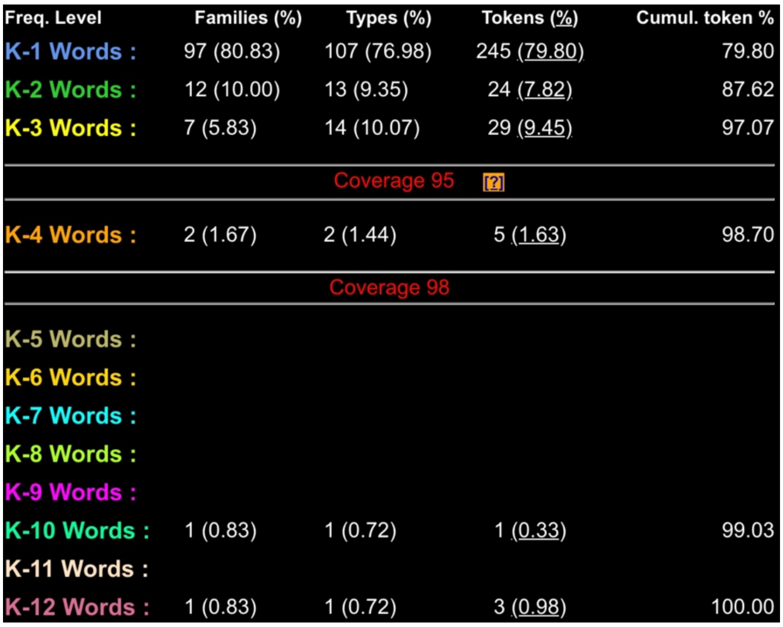
Figure 4, 5 and 6 display the lexical frequency level for CEFR A2, B1 and B2 level texts. Figure 4 presents the lexical frequency level for CEFR A2 level text. 99.07% of the words are within K-3 level, except two are from K-4 level (*tolerate, patience*) and one is from K-10 level (*dote*). But *dote on* is a phrase appears in the task which asks test takers to choose the correct answer from a multiple-choice question.



Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
K-1 Words :	123 (86.62)	142 (87.65)	302 (93.50)	93.50
K-2 Words :	11 (7.75)	11 (6.79)	13 (4.02)	97.52
		Coverage 95 ?		
K-3 Words :	5 (3.52)	5 (3.09)	5 (1.55)	99.07
		Coverage 98		
K-4 Words :	2 (1.41)	2 (1.23)	2 (0.62)	99.69
K-5 Words :				
K-6 Words :				
K-7 Words :				
K-8 Words :				
K-9 Words :				
K-10 Words :	1 (0.70)	1 (0.62)	1 (0.31)	100.00

Figure 4. Lexical frequency level for CEFR A2 level text

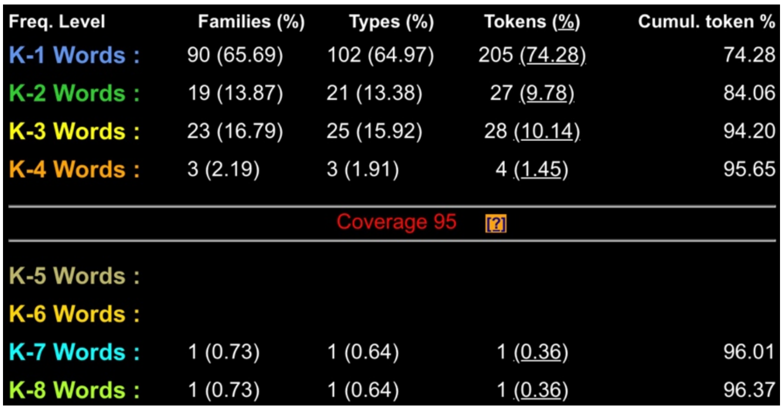
Figure 5 shows the lexical frequency level for CEFR B1 level text. 98.70% of the words are within K-4 level, except one is from K-10 level (*disloyal*) and one is from K-12 level (*Barnacles*). If test takers know the meaning of the word *loyal*, they can guess the meaning of the word *disloyal*. As for the word *Barnacles*, actually it is a task that asks test takers to guess the meaning of this word.



Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
K-1 Words :	97 (80.83)	107 (76.98)	245 (79.80)	79.80
K-2 Words :	12 (10.00)	13 (9.35)	24 (7.82)	87.62
K-3 Words :	7 (5.83)	14 (10.07)	29 (9.45)	97.07
		Coverage 95 ?		
K-4 Words :	2 (1.67)	2 (1.44)	5 (1.63)	98.70
		Coverage 98		
K-5 Words :				
K-6 Words :				
K-7 Words :				
K-8 Words :				
K-9 Words :				
K-10 Words :	1 (0.83)	1 (0.72)	1 (0.33)	99.03
K-11 Words :				
K-12 Words :	1 (0.83)	1 (0.72)	3 (0.98)	100.00

Figure 5. Lexical frequency level for CEFR B1 level text

Figure 6 presents the lexical frequency level for CEFR B2 level text. 95.65% of the words are within K-4 level, except one is from K-7 level (*grappling*) and one is from K-8 level (*pertain*). Just like mentioned above, *grappling with* is a phrase appears in the task which asks test takers to choose the correct answer from a multiple-choice question. As for the word *pertain*, it has no influence on the whole text even if the test takers do not understand this word.



Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
K-1 Words :	90 (65.69)	102 (64.97)	205 (74.28)	74.28
K-2 Words :	19 (13.87)	21 (13.38)	27 (9.78)	84.06
K-3 Words :	23 (16.79)	25 (15.92)	28 (10.14)	94.20
K-4 Words :	3 (2.19)	3 (1.91)	4 (1.45)	95.65
		Coverage 95	?	
K-5 Words :				
K-6 Words :				
K-7 Words :	1 (0.73)	1 (0.64)	1 (0.36)	96.01
K-8 Words :	1 (0.73)	1 (0.64)	1 (0.36)	96.37

Figure 6. Lexical frequency level for CEFR B2 level text

3.5.3 Statistical Package for the Social Sciences (SPSS, version 24.0)

A significance level of 0.05 ($p \leq .05$) was set in this study. SPSS 24.0 was used for descriptive statistics like mean, standard deviation, significance, correlation, etc. were computed. Also, an independent T-test was used to examine whether there was any difference on test takers' performance between CBT and PBT based on mean scores across different CEFR levels.

4. Results

4.1 Research Question 1: Is there a Difference between Test Takers' Performance on CBT and PBT?

4.1.1 Descriptive Analysis of Test Performance by Sections on CBT and PBT

Table 1 shows the overall descriptive statistics by sections. The skewness and kurtosis are both within -2 to 2, indicating all the test results are normally distributed. Thus, they can be used for further analysis. The mean of reading section is 28.49 (S.D.=5.07) on CBT (N=53) and 28.27 (S.D.=5.29) on PBT (N=236). The gap is only 0.22 which means there is no big difference between test takers' performance and test modes. The mean of listening section is 17.13 (S.D.=4.46) on CBT (N=53) and 15.36 (S.D.=5.04) on PBT (N=236). The gap is 1.77 which means test takers did better on CBT than PBT. The same outcome shows in the writing section, with the mean on CBT (N=53) is 20.36 (S.D.=7.29) and PBT (N=236) is 18.25 (S.D.=6.79) indicating the biggest gap 2.11. For the whole test, the mean of total score is 65.98 (S.D.=14.28) on CBT (N=53) and 61.88 (S.D.=15.11) on PBT (N=236). The gap is 4.10 indicating test takers did better on CBT than PBT.

Table 1. Overall descriptive statistics by sections on CBT and PBT (NCBT=53, NPBT=236)

Section	*Test mode*	*Mean*	*S.D.*	*Skewness*	*Kurtosis*
RSUM	CBT	28.49	5.07	-.62	-.25
	PBT	28.27	5.29	-.55	.04
LSUM	CBT	17.13	4.46	-.42	-.24
	PBT	15.36	5.04	-.13	-.81
WSUM	CBT	20.36	7.29	-.25	1.16
	PBT	18.25	6.79	-.06	-.11
TOTAL	CBT	65.98	14.28	-.32	-.28
	PBT	61.88	15.11	-.23	.30

Note. RSUM=Reading sum; LSUM=Listening sum; WSUM=Writing sum

4.1.2 Mean Difference of Test Scores by Sections between CBT and PBT

An independent T-test was conducted to examine test takers' performance on both CBT and PBT. Table 2 displays the mean difference of test scores by each section between CBT and PBT. A significance level of 0.05 ($p \leq .05$) was set in this study to examine whether it was statistically significant between test takers' performance and test modes.

The results show listening section is statistically significant (*p*= .01) between test scores and CBT and PBT. Therefore, the test takers performed better on the listening section on CBT (M=17.13, S.D.=4.46) than PBT (M=15.36, S.D.=5.04). Similar result was got from the writing section. The writing section is statistically significant (*p*= .05) between test scores and CBT and PBT. Therefore, the test takers performed better on the listening section on CBT (M=20.36, S.D.=7.29) than PBT (M=18.25, S.D.=6.79). But for the reading section, there is no statistically different (*p*= .78) between test scores and CBT and PBT. The same results were got from the total test. For the total test, there is no statistically different (*p*= .78) between test scores and CBT and PBT.

Table 2. Mean difference of test scores by sections between CBT and PBT (NCBT=53, NPBT=236)

Section	*Mean difference*	*Std. Error difference*	*t*	*df*	*Sig. (2-talied)*
RSUM	.22	.81	.27	287	.78
LSUM	1.77	.63	2.86	94.40	.01
WSUM	2.11	1.05	2.02	287	.05
TOTAL	4.10	1.41	1.45	287	.15

4.1.3 Inter-Rater Reliability

As mentioned before, the writing section was scored by two ELC staff. Thus, the inter-rater reliability was conducted to see whether double scoring was reliable or not. Table 3 shows that the scores by Rater 1 are slightly higher than scores by Rater 2 for both tasks. Since scores are normally distributed, further correlation analysis can be conducted.

Table 3. Descriptive analysis of scores by the two raters

Rater	*Mean*	*S.D.*	*Skewness*	*Kurtosis*
E-mail R1	2.06	1.25	-.05	-.43
E-mail R2	1.96	1.34	.12	-.88
Essay R1	13.19	4.77	-.63	1.10
Essay R2	12.43	5.90	-.16	-.80

Note. R1=Rater 1; R2=Rater 2

Table 4 shows that the e-mail scores by the two raters are strongly correlated (r = 0.90, p < 0.01) and the essay scores by the two raters are moderately correlated (r = 0.69, p < 0.01) indicating scores for the writing section are reliable.

Table 4. Correlation between scores by the two raters

	E-mail R2	*Essay R2*
E-mail R1	.896**	
Essay R1		.691**

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

4.2 Research Question 2: Is there a Difference between Test Takers' Performance on CBT and PBT across Tasks at Different CEFR Levels?

4.2.1 Descriptive Analysis of Test Performance by CEFR Levels on CBT and PBT

Table 5 presents the overall descriptive statistics by different CEFR levels. The skewness and kurtosis are both within -2 to 2, indicating all the test results are normally distributed. Thus, they can be used for further analysis. The mean score of tasks at CEFR A2 level is 18.43 (S.D.=1.40) on CBT (N=53) and 25.15 (S.D.=2.61) on PBT (N=236). The gap is 6.72 which means there is a very big difference between test takers' performance and test modes, and test takers performed better on PBT than CBT. The mean score of tasks at CEFR B1 level is 21.50 (S.D.=1.45) on CBT (N=53) and 17.36 (S.D.=1.89) on PBT (N=236). The gap is 4.14 which means test takers did better on CBT than PBT. The same outcome shows in the mean scores of tasks at CEFR B2 level, with the mean on CBT (N=53) is 25.42 (S.D.=5.11) and PBT (N=236) is 19.37 (S.D.=4.15) indicating the gap 6.05. These results are rather different from the above descriptive statistics by sections.

Table 5. Overall descriptive statistics by CEFR levels on CBT and PBT (NCBT=53, NPBT=236)

Task difficulty	*Test mode*	*M*	*S.D.*	*Skewness*	*Kurtosis*
CEFR A2 level	CBT	18.43	1.40	-.69	.38
	PBT	25.15	2.61	-1.15	1.49
CEFR B1 level	CBT	21.50	1.45	-.80	.17
	PBT	17.36	1.89	-.51	-.76
CEFR B2 level	CBT	25.42	5.11	-.16	.03
	PBT	19.37	4.15	-.06	-.12

4.2.2 Mean Difference of Test Scores by CEFR Levels between CBT and PBT

An independent T-test was conducted to examine test takers' performance on both CBT and PBT. Table 6 displays the mean difference of test scores by tasks at different CEFR levels between CBT and PBT. A significance level of 0.05 ($p \leq .05$) was set in this study to examine whether it was statistically significant between test takers' performance and test modes.

The results show the mean difference of tasks at CEFR A2 level is statistically significant (*p*= .03) between test scores and CBT and PBT. Therefore, the test takers performed better at CEFR A2 level tasks on PBT (M=25.15, S.D.=2.61) than CBT (M=18.43, S.D.=1.40). But for the mean difference of tasks at CEFR B1 level, there is no statistically different (*p*= .49) between test scores and CBT and PBT. The same results were got from the tasks at CEFR B2 level. There is no statistically different (*p*= .18) between test scores and CBT and PBT.

Table 6. Mean difference of test scores by CEFR levels between CBT and PBT (NCBT=53, NPBT=236)

Task difficulty	*Mean difference*	*Std. Error difference*	*t*	*df*	*Sig. (2-talied)*
CEFR A2 level	6.72	.19	.36	96	.03
CEFR B1 level	4.14	.29	.69	287	.49
CEFR B2 level	6.05	.33	1.36	287	.18

5. Limitations

Some limitations were found in this study. To begin with, the number of test takers took part in CBT and PBT was not equal. 236 test takers took the PBT and 56 took the CBT. It would be better if the number of the participants is equal or the gap of participants is not such big for further study. And the study sample for CBT is too few and it could be larger for the further study. Furthermore, test takers took part in only one mode. It would be better if they are able to take both modes of the test for the further study. Also, it is hard to control some of the variables, such test takers' familiarity with the computer, the speed they read on computer and paper, etc. It would be better to control some more variables for the further study. Finally, the study only focused on the language skills of reading, listening and writing. It would be better if speaking could also be taken into consideration for the further study.

6. Conclusion

This study compared the test takers' performance on both CBT and PBT across tasks with different CEFR levels. It showed that there was a difference between test takers' performance on CBT and PBT based on mean scores, and listening section was statistically significant between test scores and CBT and PBT. Therefore, the test takers performed better on the listening section on CBT than PBT. Similar results were got from the writing section. But for the reading section, there was no statistically different between test scores and CBT and PBT. The same results got from the total test, which were consistent with Anakwe's research in 2008 and Öz and Özturan's research in 2018.

In terms of CEFR levels, the results showed the mean difference of tasks at CEFR A2 level was statistically significant between test scores and CBT and PBT. Therefore, the test takers performed better at CEFR A2 level tasks on PBT than CBT. But for the mean difference of tasks at CEFR B1 level, there was no statistically different between test scores and CBT and PBT. The same results were got from the tasks at CEFR B2 level.

Also, inter-rater reliability was conducted to examine the reliability of two raters. The result showed the high correlation for the E-mail writing section and moderate correlation for the essay section. Thus, the scores of the writing section were reliable.

In summary, there was a difference between test takers' performance on different test modes across different CEFR levels, but only CEFR A2 level showed statistically difference between CBT and PBT. And since science and technology are continuously developing, it is essential for the university to consider switching the test mode from PBT to CBT.

References

Anakwe, B. (2008). Comparison of student performance in paper-based versus computer-based testing. *Journal of Education for Business, 84*(1), 13-17. https://doi.org/10.3200/JOEB.84.1.13-17

Choi, I., Kim, K. & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing, 20*(3), 295-320. https://doi.org/10.1191/0265532203lt258oa

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

Gallagher, A., Bridgeman, B. & Cahalan, C. (2000). The effect of computer-based tests on racial/ethnic, gender, and language groups. *ETS Research Report Series, 2000*(1), 1-17. https://doi.org/10.1002/j.2333-8504.2000.tb01831.x

Goldberg, A. & Pedulla, J.J. (2002). Performance differences according to test mode and computer familiarity on a practice GRE. *Educational and Psychological Measurement, 62*(6), 1053-1067. https://doi.org/10.1177/0013164402238092

Jeong, H. (2011). A comparative study of scores on computer-based tests and paper-based tests. *Behavior and Information Technology, 33*(4), 410-422. https://doi.org/10.1080/0144929X.2012.710647

Mazzeo, J. & Harvey, A.L. (1988). The equivalence of scores from automated and conventional educational and psychological tests. *A review of the literature* (ETS RR 88-21). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1988.tb00277.x

Mead, A.D. & Drasgow, F. (1993). Equivalence of computerized and paper cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449-458. https://doi.org/10.1037/0033-2909.114.3.449

Oduntan, O., Ojuawo, O. & Oduntan, E. (2015). A comparative analysis of student performance in paper pencil test (PPT) and computer-based test (CBT) examination system. *Research Journal of Educational Studies and Review, 1*(1), 24-29. http://www.pearlresearchjournals.org/journals/rjesr/archive/2015/April/pdf/Oduntan%20et%20al%20.pdf

Öz, H. & Özturan, T. (2018). Computer-based and paper-based testing: Does the test administration mode influence the reliability and validity of achievement tests? *Journal of Language and Linguistic Studies, 14*(1), 67-85. http://jlls.org/index.php/jlls/article/view/878

Parshall, C. & Kromrey, J. D. (1993). *Computer testing versus paper-and-pencil testing: An analysis of examinee characteristics associated with mode effect.* Paper presented at the Annual Meeting of the American Educational Research Association. Atlanta, GA.

Wang, S.D., Jiao, H., Young, M.J., Brooks, T. & Olson, J. (2007). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*(1), 5. https://doi.org/10.1177/0013164407305592

Copyrights