Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 13, No. 1, March 2023

ISSN 2476-5880

Delineating Discrepancies between TOEFL PBT and CBT

Ahmad Yulianto[1]*, Anastasia Pudjitriherwanti[2], Chevy Kusumah[3], Dies Oktavia[4]

Received: 28 September 2022 Accepted: 8 December 2022

Abstract
The increasing use of computer-based mode in language testing raises concern over its similarities with and differences from paper-based format. The present study aimed to delineate discrepancies between TOEFL PBT and CBT. For that objective, a quantitative method was employed to probe into scores equivalence, the performance of male-female participants, the relationship between completion time and test score, and test mode's effects on participants' performance. Totally, 124 undergraduates partook in the current research whose ages ranged from 19 – 21 years (*M* = 20, *SD* = .66). To analyze the data, MANOVA, Pearson correlation, and regression tests were run. The findings uncovered that: (1) PBT and CBT were equivalent in scores; (2) male and female's scores were not significantly different; (3) there was a moderately negative correlation between completion time and score; (4) computer familiarity, habit in using computers, and perception toward CBT did not affect performance in TOEFL. For researchers, the implication of this study concerns the interchangeability of the two-test modes. For CBT test designers, it concerns the appropriate inclusion of visuals, time related measurement, and procedures to design computer-based tests.

Keywords: CBT; computer familiarity; discrepancies; equivalent; PBT

1. Introduction

As information and communication technology is rapidly progressing, computer-based test (henceforth CBT) has become an alternative to the conventional testing mode of the paper-based test (henceforth PBT). CBT is a possible substitute for the conventional paper-based test (Ary et al., 2018; He & Tymms, 2005; Smoline, 2008; Triantafillou et al., 2008). More and more computer-based tests are employed for numerous types of testing and purposes. Thelwell (2000) and Russo (2002) exemplified that different tests like job applications, military training, TOEFL, and GRE tests have also relied on CBT. In Indonesia, a concrete example of CBT implementation is the replacement of PBT with CBT for national examination (Mangesa et al., 2021). There is a consensus in the literature that CBT is inevitably becoming part of education worldwide nowadays (Lehane et al., 2022).

[1] French Literature Department of Universitas Negeri Semarang, Email: ay@mail.unnes.ac.id
[2] French Literature Department of Universitas Negeri Semarang, Email: astaputri@mail.unnes.ac.id
[3] Japanese Education Department of Universitas Negeri Semarang, Email: chevykusumah84@mail.unnes.ac.id
[4] French Education Department of Universitas Negeri Semarang, Email: dies_oktavia@mail.unnes.ac.id

This preference is reasonable since CBT offers more convenient administration and practicalities. The efficacy and ease of CBT also promoted its utilization in the assessment programs within educational fields (Keng et al., 2008). CBT enables us to score instantaneously and get immediate response, allows an individualized examining method, improves test management, and cuts down costs (Akdemir & Oquz, 2008; Paek, 2005). Much has been said about the advantages of CBT utilization in language testing like immediate results which are easier to be analyzed (Boevé et al., 2015; Laborda & Penalver, 2018) and the opportunity for the participants to interact positively with the questions and get feedback instantly (Daniels & Gearls, 2017).

However, replacing paper tests with a computerized test format is not free from consequences. Some research discovered potential drawbacks in the use of CBT (O'Malley et al., 2005; Paek, 2005). It is empirically evidenced that distinctions exist in the replacement of PBT by CBT (Cerillo & Davis, 2004). CBT requires participants to be familiar with the basics of computer operation. To see, choose the item, and select the answer from the list of choices, participants must be able to operate the keyboard and mouse aptly. The more familiar the participants with computer operations are, the bigger their chances will be to have a better result (Mangesa et al., 2021).

Some factors are believed to affect how someone performs in CBT like his/her computer skills, attitude toward computer use, and even anxiety about using a computer in the test. Bachman and Palmer (1996) stated that using computers in an examination brought various effects on the performance of participants. Thus, people who are not adaptive tend to be avoidant or reluctant which in turn might affect their performance in tests. Provided that the main difference between PBT and CBT resides in how the test is administered, some experts argue that the two test modes may be regarded comparable (e.g., Neuman & Baydoun, 1998). Conducting tests on a computer brings a dissimilar atmosphere that may influence the way the participants perform.

While many years ago ICT illiteracy constituted a problem, nowadays as computers are extensively used this view is probably no longer true. To what degree the change of test mode from the traditional way of paper-based test to a computerized fashion may affect participants' performance is a critical question to answer and all the current assumptions about testing mode effect, be it positive or negative, badly need reviewing.

2. Review of Literature

2.1. PBT to CBT Comparison

The different performance due to the test mode in use i.e., PBT or CBT has been investigated in recent years. Some studies favored the assumption that the test-mode shift had a significant impact on participants' performance while others showed otherwise. Among those who agreed on the impact of CBT on performance in the test were Parshall and Kromrey (1993). They revealed that participants performed better in CBT than in PBT as demonstrated in their study of 1,114 participants who took the Graduate Record Examination. Mangen et al. (2013) showed the differences in the test-taker's performance in their study of 72 students. They discovered that these students obtained higher scores in CBT. Likewise, Washburn et al. (2017) examined the performance and perceptions of participants toward CBT and PBT, especially in the transition of the test method. Their findings revealed that CBT scores were greater than PBT scores.

However, Khoshsima et al. (2019) reported different results. No statistically substantial variances were found in the participants' test score of PBT and CBT. Öz and Özturan (2018) also supported this finding when they examined 97 Turkish students for paper and computer-based tests. The result revealed that the test method did not affect their performance.

2.2. Testing Mode Effect

While the exact cause of different scores in CBT and PBT is still debatable, computer familiarity, perception toward CBT, intelligence, and educational background of the participants are believed to be influential factors. These factors are crucial in some ways and may determine the results of CBT (Russel & Haney, 1997; Vispoel et al., 2001). Other scholars proposed the possible effects of intervening variables like computer familiarity (Jeong, 2014), attitude toward the use of computers (Dammas, 2016), computer aversion (Balogun & Olanrewaju, 2016), and mode preference (Boevé et al., 2015; Mizrachi, 2015) on the test scores. How the test administration influences the participants' performance is called the testing mode effect.

The testing mode effect can be negative or positive as stated by McDonald (2002). He put forward a negative example called computer aversion. It is a feeling of displeasure or discomfort experienced when someone does a test on a computer. However, the real impact of computer aversion on test takers' performance is still controversial despite the fact showing that participants who object to computer utilization usually perform poorly on CBT (Balogun & Olanrewaju, 2016).

Other studies have shown that participants have a positive view of CBT and prefer CBT when required to choose between CBT or PBT (Al-Amri, 2009). Although research evidence in the university context concludes this, Khoshsima & Toroujeni (2017) however, contended that the variables mentioned above cannot be considered as factors that affect student performance in CBT. Nowadays learners are more familiar with computers through playing games or using the internet and communicating via various types of messengers. That is why computer familiarity is probably losing its significance and relationship with CBT performance.

2.3. Characteristics of CBT

Chalhoub-Deville (as cited in Milanovic, 2001) described the characteristics of CBT as follows. First, unlike PBT which presents information via text and audio only, CBT displays multimedia features like graphics, captions, videos, or audio in the sense that the presentation is close to reality. Second, listening comprehension is carried out by displaying an image on the monitor that accompanies the audio. These features help to simulate the genuineness of the situational context so that the authenticity and validity of the test are more guaranteed and serve as a clue to the situation being discussed. CBT also brings greater flexibility in the timing of tasks for the test-takers because they can pace their work. These advantages may in turn help reduce test anxiety and frustration. Regardless of these advantages, some participants complained of eye fatigue due to prolonged exposure to computer screens (Larson, 1999).

2.4. Measuring Participants' Performance in PBT and CBT

Measuring test-takers' performance either in PBT or CBT is not easy and should be conducted carefully. Theoretically, a person who achieved a good score in PBT was supposed to have a comparable score in CBT (Sangmeister, 2017). The CBT International Guidelines state that identical tests run in two different ways should result in equal and reliable scores (International Test Commission, 2006).

To investigate the equivalence of these two test modes, distribution, rank, and score correlation should be discussed in regard to their psychometric characteristics. If the results are satisfactory, then the two tests are considered comparable (Van de Vijver & Harsveld, 1994; Wang & Kolen, 2001). Second, if the specified criteria are met, then a higher research method can be carried out, for example by equation modeling or confirmatory factor analysis. To examine how the test-takers' preferences are correlated with the scores obtained, a preference scale questionnaire or interview can be used (Al-Amri, 2009; Corlett-Rivera & Hackman, 2014; Mizrachi, 2015). Given the above-mentioned arguments, here are the aspects that the current study attempted to elucidate. Firstly, most studies only emphasized the comparison of TOEFL Total Score. The present study tried to explore score variations of each section in TOEFL PBT and CBT since it might help reveal the characteristics of these two testing modes. Secondly, male and female score comparison has indeed been discussed in some research. Males are believed to be more apt in technology-related domains like CBT, the present study tried to demystify this supposition. Thirdly, participants in CBT relatively have more freedom to manage time during the test. This privilege deserves an investigation as to its influence on the participants' performance. Finally, the previous studies mostly took place within the past few years when the exposure to computers, cell phones, and the internet was not as much as it is now. Whether participants have difficulty doing a test on a computer or not is the question that the present study attempted to tackle. Accordingly, this study aimed to address these issues:

1) Do the scores of TOEFL PBT and CBT significantly differ?
2) Do male participants outperform female participants in TOEFL CBT?
3) Does less completion time in CBT signify the participants' better performance in TOEFL?
4) Do factors like computer familiarity, computer habits, and perception toward CBT affect participants' performance in TOEFL CBT?

3. Method

3.1. Research Design

A quantitative method employing TOEFL PBT, CBT, and a questionnaire of computer familiarity, computer habit, and perception toward CBT (CFHP) was chosen for this study. This method was deemed appropriate since this study principally intended to reveal discrepancies between PBT and CBT.

3.2. Participants

Totally, 124 sophomores registered in the French Department of Universitas Negeri Semarang participated in this study. They were from the 2018 and 2019 batches with a composition of 29 male students (23.39 %) and 95 female students (76.61%). Their age ranged from 19 to 21 years (*M* = 20 and *SD* = .66). On average, all participants had received 7 years of English instruction from elementary to high school and 1 semester at the university level by the time of this study. Before the test, the participants were given a consent form to sign.

3.3. Instrumentation

Three instruments were used in this research:

1) TOEFL Preparation Course by Deborah Phillips, 2001 Edition.
2) TOEFL CBT of similar material developed on http://www.ujian.unnes.ac.id by the Language Centre of Universitas Negeri Semarang to accommodate internal testing.
   Both versions were presented in multiple-choice questions format and consisted of 3 units: listening comprehension (50 problems), structure & written expression (40 problems), and reading comprehension (50 problems). Reliability tests showed that all cases were valid with no case excluded (N=124/100%) for TOEFL PBT and CBT. Cronbach's Alpha showed a strong correlation coefficient value of .96. To examine construct validity of these two tests, exploratory factor analysis (EFA) was employed. This analysis was performed to find out whether the initial factors (items) of 140 represented the subscale factors in listening, structure, and reading comprehension. Initial test of Kaiser-Mayer-Olkin (KMO) showed that the value of PBT was .631 while that of CBT was .672; both met the minimum criteria for factor analysis. Based on the data analysis, there were 52 factors for the TOEFL PBT and 56 factors for CBT with Eigenvalues higher than 1. Both were fewer than the number of factors before the extraction (140) but were considered representative of the subscale factors contained in the tests. This factor structure provides evidence for the construct validity of the tests.
3) Questionnaire on computer familiarity, habit, and perception toward CBT designed from CAAFI Index by Schulenberg and Melton (2008). This questionnaire (henceforth CFHP) consisted of 27 items and covered 3 aspects i.e., computer familiarity (items 1 – 13), the habit of computer use (items 14 – 21), and perception toward the computer-based mode of testing (items 22 – 27). The validity test with a significance of 5% indicated that no item had a value less than .18, the highest value of .70, and the lowest value of .18. Reliability test showed that all cases were valid with no case excluded (N=124/100%). Cronbach's Alpha showed a strong correlation coefficient value of .77. The CFHP questionnaire was valid and reliable; therefore, it could be used in this study.

3.4. Procedure

Three-stage testing was run for the current study. Initially, participants took TOEFL PBT at the end of the Odd Semester 2020 – 2021. In the second stage, they took TOEFL CBT at http://www.ujian.unnes.ac.id. This test was administered online at the beginning of the Even

Semester 2020 – 2021. A direction was presented to the test-takers ahead of time concerning the know-how and their response confidentiality. To ensure that participants are familiar with the CBT procedure, an instruction was made available in a YouTube video. This procedure complied with Akdemir's and Oguz's opinion (2008) that scrutinized possible distinctions of these two test modes. This was also in line with Momeni (2022) who stated that before online assessment, learners should be given complete and thorough instructions. After that, participants were required to fill out the questionnaire distributed on Google Forms. They selected one of the options either *Strongly Disagree*, *Disagree*, *Neutral*, *Agree*, or *Strongly Agree*. These options were then converted to Likert Scale 1 – 5 for statistical analysis.

3.5. Data Collection

TOEFL PBT and CBT were administered at the college in which the participants were pursuing their degree. Collection and checking of PBT data were done manually while CBT data were derived from http://www.ujian.unnes.ac.id. The scoring was done categorically (an incorrect answer is worth 0 while a correct answer is worth 1 point). Both CBT and PBT raw scores were then manually converted using TOEFL conversion table. As to the CFHP questionnaire, the data were collected from the respondents' responses stored on google drive.

3.6. Data Analysis

To analyze the data, these methods were employed. Firstly, descriptive statistics and Multivariate Analysis of Variance (MANOVA) were run to check whether different scores occurred in the TOEFL PBT and CBT and discover whether scores were different between males and females. Afterward, a linear combination of the three measures i.e., listening, structure, and reading of PBT and CBT was computed. Pearson *r* correlation and regression analyses were then run to discover if there was a relationship between completion time and CBT scores. Finally, regression analysis was conducted to discover if the level of computer familiarity, habits in using computers, and perception toward CBT affected CBT scores. All the data analyses were done on SPSS 25.

4. Results

The result of the preliminary analysis is presented in the descriptive statistics below.

4.1. PBT vs. CBT Score Comparison

Table 1.
Descriptive Statistics of TOEFL PBT and CBT Scores

	Test Mode	Gender	*M*	*SD*	*N*
Listening	PBT	Male	46.52	7.61	29
		Female	47.68	7.44	95
		Total	47.41	7.46	124
	CBT	Male	45.79	7.42	29
		Female	47.26	7.84	95
		Total	46.92	7.74	124
Structure	PBT	Male	45.45	7.75	29
		Female	46.33	7.34	95
		Total	46.12	7.41	124
	CBT	Male	45.62	8.10	29
		Female	46.28	7.65	95
		Total	46.13	7.73	124
Reading	PBT	Male	45.76	8.81	29
		Female	48.42	7.33	95
		Total	47.80	7.75	124
	CBT	Male	46.59	9.78	29
		Female	49.15	6.97	95
		Total	48.55	7.75	124
Overall	PBT	Male	454.31	61.73	29
		Female	472.93	59.43	95
	Total		468.57	60.25	124
	CBT	Male	462.72	62.69	29
		Female	475.39	60.88	95
	Total		472.43	61.29	124

Table 1 shows that the total score of PBT (*M* = 468.57, *SD* = 60.25) was lower than that of CBT (*M* = 472.43, *SD* = 61.29). A comparison of subsection scores demonstrated that PBT was higher in listening than CBT but lower in structure and reading. In gender comparison, female participants (*M* = 472.93, *SD* = 59.43) outscored male participants (*M* = 454.31, *SD* = 61.73) in PBT total score. Female participants (*M* = 475.39, *SD* = 60.88) also outperformed male participants (*M* = 462.72, *SD* = 62.69) in CBT total score. Female participants' scores in all three sections of CBT i.e., listening, structure, and reading were also greater than male participants' scores. To further investigate these discrepancies, a MANOVA test was run. No considerable abnormalities were found on the homogeneity of variance-covariance matrices and normality assumptions.

Table 2.
Results of One-way MANOVA

Effect	Wilks' Λ Value	*F*	Hypothesis *df*	Error *df*	*P*	Partial Eta Squared
Test Mode	.99	.52	3	243	.67	.01
Gender	.97	2.20	3	243	.09	.03

Table 2 shows that based on test mode, no significant difference of test scores was found, *F* (3, 243) = .52, *p* = .67, Wilk's Λ = .99, $\eta^2{}_p$ = .01. So, the participants' performance in TOEFL was not dependent on the test mode taken. Based on gender, the outcome indicated that the test scores were not statistically different with *F*(3, 243) = 2.20, *p* = .09; Wilk's Λ = .97, $\eta^2{}_{p\,=}$ .03. Thus, we can state that the participants' performance in TOEFL did not depend on gender. To probe the impact of the testing mode and gender on each section, tests of between-subject effects were performed.

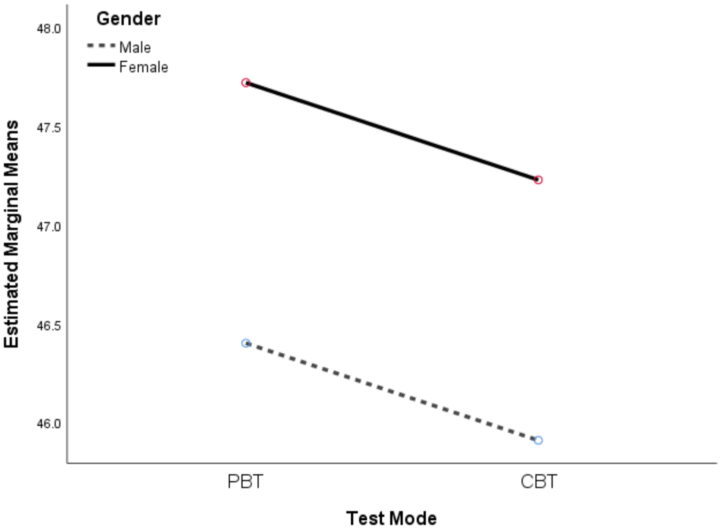
Table 3.
Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Test Mode	Listening	15.00	1	15.00	.26	.61
	Structure	.00	1	.00	.00	.99
	Reading	34.88	1	34.88	.59	.44
Gender	Listening	77.25	1	77.25	1.34	.25
	Structure	26.40	1	26.40	.46	.50
	Reading	303.12	1	303.12	5.13	.02
Error	Listening	14139.97	245	57.71		
	Structure	14080.72	245	57.47		
	Reading	14467.55	245	59.05		

Table 3 displays that test mode did not significantly affect listening score, *F*(1, 245) = .26, *p* = .61. Also, it did not have an impact on structure score, *F*(1, 245) = .00, *p* = .99. Similarly, it had no significant effect on reading score, *F*(1, 245) = .59, *p* = .44. Gender didn't seem to have any significant effect on listening score, *F* (1, 245) = 1.34, *p* = .25. It also had no significant impact on structure score, *F* (1, 245) = .46, *p* = .50 as well as on reading score, *F* (1, 245) = 5.13, *p* = .03. So, it can be stated that no significant discrepancies were found on the listening, structure, and reading scores across gender. The following figures depict the estimated marginal means for each section while showing the interaction between test mode and gender.

Figure 1.

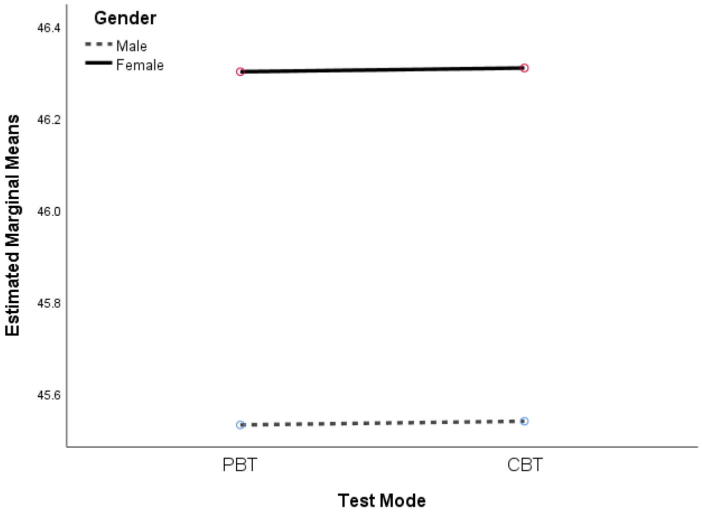
Interactions between Male and Female Participants for Listening



Note: Interaction in listening scores show the same tendency for both male and female participants where PBT score was a bit greater than CBT score although insignificant in pairwise comparison test (*MD* = 6.17, *SE* = .28, *p* = .61).

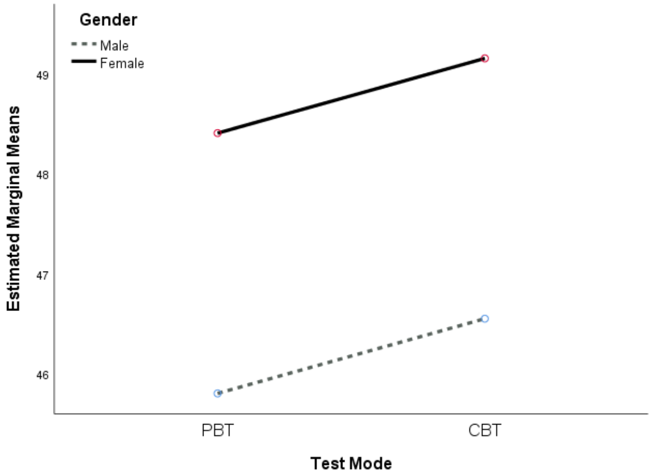
Figure 2.

Interactions between Male and Female Participants for Structure



Note: Male and female participants performed almost similarly in the structure section of PBT and CBT. CBT structure revealed a slightly higher score than PBT structure though not significant (*MD* = 6.25, *SE* = .27, *p* = .99).

Figure 3.
Interactions between Male and Female Participants for Reading



Note: Predicted score for both male and female participants was greater in CBT reading. However, this difference was found insignificant in the pairwise comparison test (*MD* = 6.30, *SE* = .28, *p* = .45).

Mean absolute deviations of listening, structure, and reading suggested that no significant variances existed either in male or female participants. The scores pattern for each section i.e., listening, structure, and reading across gender were not much different either. The data were then examined for a linear relationship between CBT score and completion time.

4.2. Results of Correlation Analyses for Completion Time and CBT Score

The following table presents correlations between completion time and CBT scores.

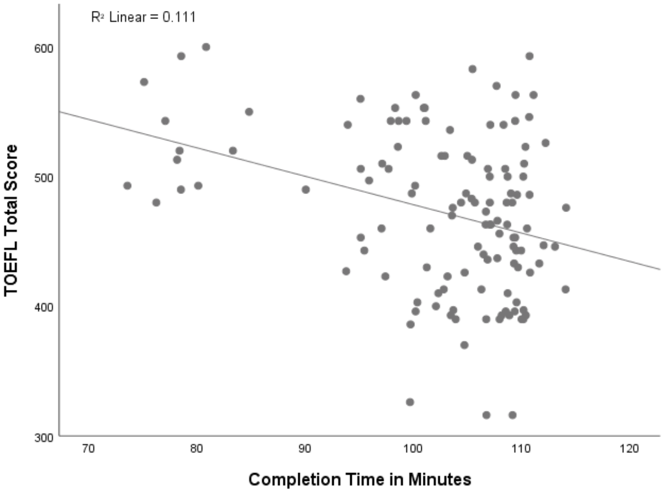
Table 4.
Pearson Correlation of Completion Time and CBT Score

	Completion time	
	r	Sig. (2-tailed)
Listening Score	-.08	.41
Structure Score	-.31**	.00
Reading Score	-.44**	.00
CBT Total Score	-.33**	.00

** Significant at the 0.01 level (2-tailed).

Table 4 suggests that completion time and listening score were negatively correlated, *r*(122) = -.08, *p* < .05. A negative correlation at a moderate level was also found in the interaction between completion time and structure score, *r*(122) = -.31, *p* < .01 as well as between completion time and reading score, *r*(122) = -.44, *p* < .01. There was also a negatively linear relationship between completion time and TOEFL total score at a medium level, *r*(122) = -.33, *p* < .01. So, when the completion time increased, the TOEFL total score decreased and otherwise. This correlation appears more clearly in the scatterplot below.

Figure 4.
Association between Completion Time and TOEFL Total Score



Note: Each dot represents an individual participant. Scores for TOEFL were obtained in CBT test mode. Completion time refers to the time each participant needed to complete the test.

Figure 4 shows that the data consist of a smaller and a bigger set. The smaller set contains a few participants who achieved higher scores in less time. The bigger group contains participants who achieved higher scores in a relatively longer time and also those who spent much time but made lower scores instead. This result indicated that the change in completion time was inversely proportional to the TOEFL total score. Regression analysis was then run to examine if completion time could predict CBT Scores.

Table 5.
Regression Outcome

Predictor	*R*	*R²*	*R²* Change	*F*	*df1*	*df2*	Sig.
Model 1	.08	.06	.06	.68	1	122	.41
Model 2	.31	.10	.10	13.07	1	122	.00
Model 3	.44	.19	.19	28.59	1	122	.00
Model 4	.33	.11	.11	15.24	1	122	.00

Model 1: Completion time on Listening Score; Model 2: Completion time on Structure Score; Model 3: Completion time on Reading Score; Model 4: Completion time on CBT Total Score

As reported in table 5, multivariate analyses were computed resulting in the extrapolative power of completion-time to score. Model 1 shows that completion time contributed an insignificant percentage of the variance (6%) of the listening scores while the other 94% was determined by other variables (*F*(1, 122) = .68, *p* > .05). Model 2 suggests that completion time explained 10% of the variance in structure score while the other 90% were determined by other variables (*F*(1, 122) = 13.07, *p* < .01). Model 3 indicates that completion time explained 19% of the variance in reading score while the other 81% were determined by other variables (*F*(1, 122) = 28.59, *p* < .01). Model 4 indicates that completion time explained only 11% of the variance in CBT total score while the other 89% were determined by other variables (*F*(1, 122) = 15.24, *p* < .01).

4.3. Results of Computer Familiarity, Habit, and Perception toward CBT (CFHP) Questionnaire

The CFHP questionnaire result is displayed in the table below.

Table 6.
Respondents' Computer Familiarity, Habit, and Perception toward CBT

Category	Item	Frequency					Likert Score	
		Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	*M*	*SD*
Computer Familiarity	Q1	0	0	9 (7.3 %)	66 (53.2 %)	49 (39.5 %)	4.32	0.61
	Q2	0	5 (4 %)	63 (50.8 %)	45 (36.3 %)	11 (8.9 %)	3.50	0.72
	Q3	0	0	22 (17.7 %)	57 (46 %)	45 (36.3 %)	4.19	0.71
	Q4	1 (0.8 %)	18 (14.5 %)	62 (50 %)	32 (25.8 %)	11 (8.9 %)	3.27	0.85
	Q5	0	10 (8.1 %)	59 (47.6 %)	44 (35.5 %)	11 (8.9 %)	3.45	0.77
	Q6	5 (4 %)	21 (16.9 %)	54 (43.5 %)	37 (29.8 %)	7 (5.6 %)	3.16	0.91
	Q7	0	3 (2.4 %)	29 (23.4 %)	52 (41.9 %)	40 (32.3 %)	4.04	0.81
	Q8	4 (3.2 %)	13 (10.5 %)	38 (30.6 %)	32 (25.8 %)	37 (29.8 %)	3.69	1.11
	Q9	1 (0.8 %)	14 (11.3 %)	31 (25 %)	46 (37.1 %)	32 (25.8 %)	3.76	0.99
	Q10	1 (0.8 %)	14 (11.3 %)	67 (54 %)	30 (24.2 %)	12 (9.7 %)	3.31	0.83
	Q11	3 (2.4 %)	20 (16.1 %)	56 (45.2 %)	34 (27.4 %)	11 (8.9 %)	3.24	0.91
	Q12	9 (7.3 %)	22 (17.7 %)	64 (51.6 %)	20 (16.1 %)	9 (7.3 %)	2.98	0.96
	Q13	10 (8.1 %)	40 (32.3 %)	58 (46.8 %)	13 (10.5 %)	3 (2.4 %)	2.67	0.86
Average							3.51	0.85
Computer Habit	Q14	0	1 (0.8 %)	8 (6.5 %)	66 (53.2 %)	49 (39.5 %)	4.31	0.63
	Q15	0	17 (13.7 %)	51 (41.1 %)	45 (36.3 %)	11 (8.9 %)	3.40	0.84
	Q16	0	3 (2.4 %)	19 (15.3 %)	57 (46 %)	45 (36.3 %)	4.16	0.77
	Q17	0	1 (0.8 %)	19 (15.3 %)	30 (24.2 %)	74 (59.7 %)	4.43	0.78
	Q18	11 (8.9 %)	45 (36.3 %)	38 (30.6 %)	24 (19.4 %)	6 (4.8 %)	2.75	1.03
	Q19	0	7 (5.6 %)	25 (20.2 %)	40 (32.3 %)	52 (41.9 %)	4.10	0.92
	Q20	1 (0.8 %)	43 (34.7 %)	48 (38.7 %)	22 (17.7 %)	10 (8.1 %)	2.98	0.94
	Q21	1 (0.8 %)	9 (7.3 %)	32 (25.8 %)	46 (37.1 %)	36 (29 %)	3.86	0.95
Average							3.75	0.86
Perception on CBT	Q22	3 (2.4 %)	22 (17.7 %)	71 (57.3 %)	18 (14.5 %)	10 (8.1 %)	3.08	0.86
	Q23	2 (1.6 %)	8 (6.5 %)	70 (56.5 %)	33 (26.6 %)	11 (8.9 %)	3.35	0.80
	Q24	3 (2.4 %)	10 (8.1 %)	89 (71.8 %)	21 (16.9 %)	1 (0.8 %)	3.06	0.62

	Q25	2 (1.6 %)	3 (2.4 %)	47 (37.9 %)	51 (41.1 %)	21 (16.9 %)	3.69	0.84
	Q26	28 (22.6 %)	23 (18.5 %)	36 (29 %)	20 (16.1 %)	17 (13.7 %)	2.80	1.33
	Q27	1 (0.8 %)	0	14 (11.3 %)	68 (54.8 %)	41 (33.1 %)	4.19	0.70
Average							3.36	0.86

The score at the computer familiarity subscale level (*M* = 3.51, *SD* = .85) indicated that most participants claimed to be fairly familiar with computer operations. It was confirmed by most preferences *Neutral* and *Agree*. However, a large variation of agreement level appeared in each aspect of computer familiarity. Only Q1, Q2, Q4, Q10, and Q12 contributed a percentage above 50%. The score at the computer habit subscale level (*M* = 3.75, *SD* = .86) indicated that participants held a moderately positive habit of computer use. Although this subscale showed a higher mean, there was only one aspect namely Q1 where the participants contributed more than half (53.2%), showing a preference for *Agree*. The subscale score for perception toward CBT (*M* = 3.36, *SD* = .86) indicated that the participants had a moderate perception toward CBT. It was confirmed with their bigger tendencies (approximately 60%) for *Neutral* (Q22, Q23, Q24) and only one for *Agree* (Q27).

4.4. Relationship between CFHP Questionnaire and CBT Score

These regression models were calculated to predict the power of the CFHP index on the CBT score.

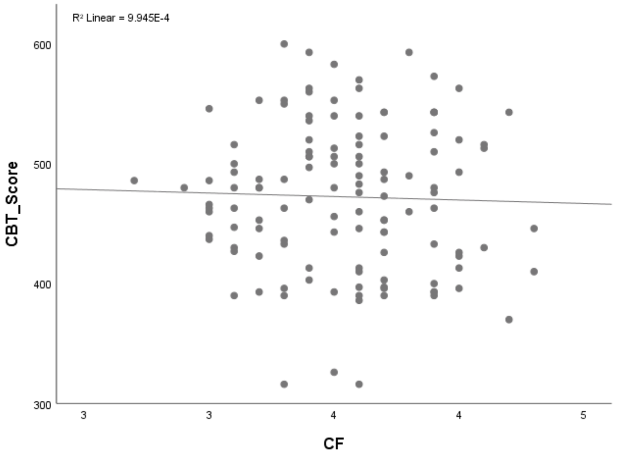
Table 7.
Regression Analysis

Predictor	*R*	R^2	R^2 Change	*F*	*df1*	*df2*	Sig.
Model 1	.041[a]	.002	.002	.206	1	122	.651
Model 2	.096[a]	.009	.009	1.135	1	122	.289
Model 3	.147[a]	.022	.022	2.696	1	122	.103
Model 4	.032[a]	.001	.001	.121	1	122	.728

CBT. Model 1: Computer Familiarity on CBT Score, Model 2: Computer Habit on CBT Score, Model 3: Perception toward computer-based test on CBT Score, Model 4: CFHP Total Score on CBT Score.

Model 1 shows that computer familiarity contributed an insignificant percentage of the variance (2%) in CBT score while the other 98% were determined by other variables (*F*(1, 122) = .206, *p* > 05) (Table 7). Model 2 shows that computer habits explained an insignificant percentage of the variance (9%) in the CBT scores while the other 91% were determined by other variables (*F*(1, 122) = 1.135, *p* > 05). Model 3 shows that perception explained a bigger percentage of the variance (22%) in CBT scores while the other 88% was determined by other variables (*F*(1, 122) = 2.696, *p* > 05). Model 4 shows that CFHP index contributed an insignificant percentage of the variance (1%) in CBT score while the other 99% were determined by other variables (*F*(1, 122) = .121, *p* > 05). This correlation appears more clearly in the scatterplot below.

Figure 5.
Association between CFHP Questionnaire and CBT Score



Note: Each dot represents an individual participant. Scores for TOEFL were obtained in CBT test mode. CF represents the participants' response to the questionnaire on computer familiarity, habit and perception toward CBT. The data points were relatively spread out, indicating no strong trend or correlation to the data.

5. Discussions

Research on PBT and CBT comparison has so far been inconclusive, mostly based on mean differences of the total score only. The present study sought to delineate the discrepancies between the two test modes more completely by examining more variables through various statistical methods.

5.1. Equivalence of PBT and CBT Scores (RQ1)

The debate over whether these two testing modes would result in equivalent or different scores is made clearer by our findings. The mean comparison between PBT and CBT did not reveal any significant differences in listening, structure, reading, and total test scores. Minor margins in the section-to-section comparison (0.49 in listening, 0.01 in structure, and 0.75 in reading) as well as in the total score (3.86) indicated that these two test modes were equal. MANOVA analysis subsequently confirmed that no discernible effect of the test mode was found, *F* (3, 243) = .52, p = .67, Wilk's Λ = .99, η2p = .01. Neither paper-based nor computer-based formats significantly contributed to the participants' performance. Thus, if the same individuals take TOEFL PBT and CBT consecutively, chances are greater that they will get approximately similar scores. These results are consistent with preceding research that mostly discovered no substantial discrepancies in the participants' total scores of PBT and CBT (Puhan, Boughton, & Kim, 2007; Wise & Plake, 1989).

Despite the equivalence mentioned above, the interaction between the variables in the two test modes revealed more to discuss. In listening, PBT score which was greater than CBT indicated that the input modalities existing in CBT like cues, pictures, or visuals did not necessarily enhance the participants' comprehension. As Pusey (2020) reiterated in his study, participants' scores in a video-equipped test did not differ significantly from their scores on the test with audio only. The result also agrees with previous studies stating that visual inputs did not necessarily help the listeners to better understand the intended message because visuals can sometimes be distracting (Ockey, 2007; Wagner, 2007). However, this is in contrast with Plass and Jones' statement (2005) that pictures and videos led to better comprehension instead.

Additionally, it is interesting to observe the interaction of the test modes in the structure section. A nearly perfect equivalence of scores between PBT (*M* = 46.12, *SD* = 7.41) and CBT (*M* = 46.13, *SD* = 7.73) suggests that performance in structure was not dependent on the test format. Structure is related to logical reasoning and grammatical rules understanding. In the meantime, in listening and reading, test-takers work based on outside stimuli. Should the test mode truly affect the score, then it is most likely to occur in listening and reading rather than in structure. The findings in reading where the CBT score was greater than PBT supported this assumption. The features available in CBT helped the test-takers in one way or another to perform better and achieve higher scores.

It is worth noting, nonetheless, that some participants were higher in total scores but were lower in one section compared to the others. It suggests that the participants' capacity in TOEFL was not equally distributed per section. Some participants were strong in listening but not good at structure and reading. Some others may be good in structure but weak in listening. Yet, some individuals may have an equal capacity in all sections.

5.2. Score Comparison of Male vs. Female Participants (RQ2)

Descriptive statistics showed that for both test modes and almost in all sections, female participants scored a bit higher than male ones although the difference was insignificant. MANOVA analysis proved that there were not any significant differences in listening, structure, reading, and total scores between male and female participants. The multivariate test result was not significant as well for gender, *F*(3, 243) = 2.20, *p* = .09; Wilk's Λ = .97, η2p = .03, indicating that there was no distinction in the level of TOEFL capacity between male and female participants. The tests of between-subjects effects also demonstrated that gender did not determine listening, structure, reading, and the total test scores. The estimated marginal means further confirmed female participants' prowess over male participants in the three sections and the total scores both for PBT and CBT; yet overall, this was not significant. It is contrary to the preceding investigations which showed that male participants scored higher than female participants in CBT (Ebrahemi & Toroujeni, 2019; Halldórsson et al., 2009; Martin & Binkley, 2009; Sørensen & Andersen, 2009; Crusoe, 2005).

Whereas female participants' scores were relatively clustered, those of male participants were more dispersed as shown by greater *SD*, indicative of an inconsistent distribution. It implies that there was not any immense gap in competencies of the female participants like what occurred in the male group. To some extent, female participants performed better both on the basis of test mode (PBT and CBT) and on the basis of sections (listening, structure, and reading). This favors the previous studies by Hyde and Linn (1988) as well as the Educational Testing Service (2007) insisting that females were slightly more advantaged in listening, reading, speaking, and writing. Again, further studies are required to confirm whether test mode is a key factor in performance on TOEFL across gender.

5.3. Correlation between Completion Time and CBT Score (RQ3)

Pearson correlations and four regression models showed a linear yet negative correlation of completion time to listening, structure, reading, and CBT total scores. The *R* values ranged from -.08 to -.33, indicating weak to moderate correlations. So, scores tended to increase when completion time decreased and otherwise. In other words, the participants who completed the TOEFL CBT in a shorter time were inclined to have higher scores. On the contrary, those who scored lower needed a longer time to complete the test. Some participants made high scores and finished the test in a relatively shorter time but some others achieved high scores in a longer time.

The result suggests that completion time could not be a key indicator of performance for the total score or subsection score in TOEFL CBT. Completion time could not predict the participants' performance in listening, structure, and reading as well as in the total score. Those who finished earlier either in listening, structure, reading, or even the whole test would not necessarily obtain good scores. This finding was, however, opposite to the study of Rafaeli and Tractinsky (2007). They maintained that there was a tight, perfect correlation of time and correctness in the computer-based test.

5.4. Effect of Computer Familiarity, Habit, and Perception toward CBT on Performance in TOEFL (RQ4)

Both item-per-item and subscales scores disclosed a modest correlation linking the CFHP index to CBT score. The score at the computer familiarity subscale (*M* = 3.51, *SD* = .85) indicated that most participants claimed to be fairly familiar with computer operations. The score at the computer habit subscale (*M* = 3.75, *SD* = .86) suggested that participants held a moderately positive habit of computer use. The subscale score for perception toward CBT (*M* = 3.36, *SD* = .86) showed that the participants had a positive perception toward CBT. In the meantime, the result of regression analysis revealed that neither at the subscale nor at the overall level did computer familiarity, the habit of using the computer, and perception toward CBT contribute significantly to the participants' performance in CBT. This result supports Yu's and Iwashita's findings (2021).

Furthermore, MANOVA analysis confirmed the absence of significantly different performances in the two types of tests. Concerning computer familiarity, it can be said that nearly everyone was familiar with using a computer; so, the traditional belief on computer familiarity's influence on performance needs questioning. Consequently, the question whether computer familiarity is related to CBT TOEFL or not no longer holds true as today's learners are more familiar with a computer through games or internet browsing and communication via different kinds of messengers. Computer familiarity is losing its importance and relationship with CBT performance. Chan (2018) and Jamieson (2005) maintained this opinion. One conceivable reason for the equivalency pertaining to PBT and CBT is people’s knowledgeability in computer operation and the like. Besides, the increasing use of computers in educational settings has augmented students’ acquaintance in computer-based tests. Scholastic tasks which tend to be ICT-based have likely influenced students’ achievement in the test as proposed by Chan et al. (2018).

6. Conclusion

This study was conducted to augment our understanding of the discrepancies between TOEFL PBT and CBT. With the intent of having a more accurate examination, mean scores comparison, the performance of male and female participants, the relationship between completion time and total score, and the effect of CFHP index on TOEFL CBT score were explored using various statistical methods. In conclusion, test mode either PBT or CBT did not yield significantly different scores in TOEFL. It was truly the participants’ competence that mattered. Then, no significant difference was discovered in the performance comparison of male and female participants. Furthermore, there was only a moderate correlation between completion time and CBT score, and that completion time could not predict the participants’ performance. And finally, participants’ index of computer familiarity, the habit of computer use, and perception toward CBT were at a moderate level, and it had no critical impact on the participants’ scores in CBT.

This study was subject to several limitations. First, it was done to Indonesian students of Universitas Negeri Semarang. Hence, the findings may be appropriate only in the university or else with comparable conditions. If applied to different participants, it might yield a different result. Also, the number of male participants accounted for 23 % only of the sample because the majority of students in the Foreign Language Department were female. Lastly, the CBT version in this study was the self-modified format of the PBT material and was made for internal use only, not the standardized computer-based test issued by official bodies. For future research, it is highly recommended that similar tests be conducted on an internationally accepted TOEFL CBT format from ETS or Pearsons to have greater accuracy of CBT performance. Not less importantly, equation modeling or confirmatory factor analysis should be used to have an in-depth comparison of TOEFL PBT and CBT scores.

The implications of this study concern especially researchers and test developers. Firstly, our findings have seemingly eliminated the doubt over TOEFL PBT and CBT equivalence and the impact of the so-called computer familiarity on CBT performance. Thus, it is time to quit the debate and start figuring out ways to improve CBT delivery so that it can measure test-takers’ performance in TOEFL accurately. Secondly, the replacement of PBT with CBT is almost inevitable in the coming years and the future test-takers are typically digital savvies who are exposed to media filled with visual cues. So, when defining and refining CBT test constructs, test developers should incorporate visuals, time-related measures, and procedures. And all should be done with utmost care so that these variables become useful instead of being disadvantageous for test-takers.

Declaration of Conflicting Interest

The authors hereby state that there are not any conflicts of interest in this study and they will be responsible for any damage or loss inflicted.

Funding

This study was funded by DIPA of Universitas Negeri Semarang Scheme No. 382.26.4/UN37/PPK.3.1/2021/dated 26 April 2021.

References

Akdemir, O., & Oguz, A. (2008). Computer-based testing: An alternative for the assessment of Turkish undergraduate students. *Computers & Education*, *51*(3), 1198–1204. https://doi.org/10.1016/j.compedu.2007.11.007

Al-Amri, S. (2009). *Computer-based testing vs. paper-based testing: Establishing the comparability of reading tests through the revolution of a new comparability model in a Saudi EFL context*. Unpublished Doctor of Philosophy in Linguistics thesis. Colchester: University of Essex. https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.496244

Ary, D., Jacobs, L., Irvine, C., & Walker, D. (2018). *Introduction to research in education*. Cengage Learning.

Bachman, L.F. & Palmer, A.S. (1996). *Language testing in practice*. New York: Oxford University Press.

Balogun, A. G., & Olanrewaju, A. S. (2016). Role of computer self-efficacy and gender in computer-based test anxiety among undergraduates in Nigeria. *Psychological Thought*, *9*(1), 58-66. https://psyct.swu.bg/index.php/psyct/article/view/160

Boevé, A. J., Meijer, R. R., Albers, C. J., Beetsma, Y., & Bosker, R. J. (2015). Introducing computer-based testing in high-stakes exams in higher education: Results of a field experiment. *PLOS ONE*, *10*(12). https://doi.org/10.1371/journal.pone.0143616

Cerillo, T., & Davis, J. (2004). Comparison of paper-based and computer-based administrations of high-stakes, high-school graduation tests. *Annual Meeting of the American Education Research Association*. San Diego.

Chalhoub-Deville, M. (Ed.). (2001). Issues in computer-adaptive testing of reading proficiency. In Milanovic, M (Eds.), *Studies in Language Testing*, *10*. Cambridge: University of Cambridge Press, 6–16.

Chan, S., Bax, S., & Weir, C. (2018). Researching the comparability of paper-based and computer-based delivery in a high-stake writing test. *Assessing Writing*, *36*, 32–48. https://doi.org/10.1016/j.asw.2018.03.008

Corlett-Rivera, K., & Hackman, T. (2014). E-book usage and attitudes in the humanities, social sciences, and education. *Portal: Libraries and the Academy*, *14*(2), 255-286. https://doi.org/10.13016/M2H91S

Crusoe, D. (2005). *A discussion of gender diversity in computer-based assessment*. Retrieved on September 29, 2021, from https://files.eric.ed.gov/fulltext/ED544707.pdf

Dammas, A. H. (2016). Investigate students' attitudes toward the computer-based test (CBT) in chemistry courses. *Archives of Business Research*, *4*(6), 58-71. https://doi.org/10.14738/abr.46.2325

Daniels, L. M., & Gierl, M. J. (2017). The impact of immediate test score reporting on university students' achievement emotions in the context of computer-based multiple-choice exams. *Learning and Instruction*, *52*, 27-35. http://dx.doi.org/10.1016/j.learninstruc.2017.04.001

Ebrahemi, M.R., & Toroujeni, M.H. (2019). Score equivalence, gender difference, and testing mode preference in a comparative study between computer-based testing and paper-based testing. *International Journal of Emerging Technologies in Learning*, *14*(7), 129-143. https://doi.org/10.3991/ijet.v14i07.10175

Educational Testing Service. (2007). *Test and score data summary for TOEFL Internet-based test: September 2005-December 2006 test data*. Princeton, NJ. Retrieved on November 7, 2021, from https://www.ets.org/Media/Research/pdf/TOEFL-SUM-0506-iBT.pdf

International Test Commission. (2006). International guidelines on computer-based and internet-delivered testing. *International Journal of Testing*, 6(2), 143-171. https://doi.org/10.1207/s15327574ijt0602_4

Halldórsson, A.M., McKelvie, P., & Björnsson, J.K. (2009). Are Icelandic boys really better on computerized tests than conventional ones? Interaction between gender, test modality, and test performance. In F. Scheuermann & J. Bjornson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing*. Luxembourg: Office for Official Publications of the European Communities. Retrieved on December 2, 2021, from https://publications.jrc.ec.europa.eu/repository/handle/JRC49408

Hyde, J. S., Linn, M. C., & Masters, J. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, *104*(1), 53-69. https://psycnet.apa.org/doi/10.1037/0033-2909.104.1.53

Jamieson, J. (2005). Trends in computer-based second language assessment. *Annual Review of Applied Linguistics*, *25*, 228–242. https://doi.org/10.1017/s0267190505000127

Jeong, H. (2014). A comparative study of scores on computer-based tests and paper-based tests. *Behavior & Information Technology*, *33*(4), 410-422. https://doi.org/10.1080/0144929X.2012.710647

Keng, L., McClarty, K.L., & Davis, L.L. (2008). Item-level comparative analysis of online and paper administrations of the Texas assessment of knowledge and skill. *Applied Measurement in Education*, *21*(3), 207–226. https://doi.org/10.1080/08957340802161774

Khoshsima, H., & Hashemi Toroujeni, S. M. (2017). Comparability of computer-based testing and paper-based testing: Testing mode effect, testing mode order, computer attitudes, and testing mode preference. *International Journal of Computer* (IJC), *24*(1), 80-99. Retrieved on February 2, 2022, from https://www.researchgate.net/publication/313820965

Khoshsima, H., Hashemi Toroujeni, S. M., Thompson, N., & Ebrahimi, M.R. (2019). Computer-based (CBT) vs. paper-based (PBT) testing: Mode effect, relationship between computer familiarity, attitudes, aversion and mode preference with CBT test scores in an Asian private

EFL context. *Teaching English with Technology, 19*(1), 86-101. Retrieved on February 2, 2022, from https://files.eric.ed.gov/fulltext/EJ1204641.pdf

Laborda, J.G., & Penalver, E.A. (2018). Constraining issues in face-to-face and internet-based language testing. *Journal for Educators, Teachers, and Trainers, 9*(2), 47-56. https://jett.labosfor.com/index.php/jett/article/view/488/372

Larson, J. (1999). Considerations for testing reading proficiency via computer-adaptive testing. In Chalhoub-Deville, M (Ed.), *Studies in language testing*, *10*. *Issues in computer-adaptive testing of reading proficiency*. Cambridge: University of Cambridge Press, 71–90.

Lehane, P., Scully, D., & O'Leary, M. (2022). Time to figure out what to do: understanding the nature of Irish post-primary students' interactions with computer-based exams (CBEs) that use multimedia stimuli, *Irish Educational Studies*, *41*(1), 5-25, https://doi.org/10.1080/03323315.2021.2022517

Mangen, A., Bente, R. W., & Kolbjørn, B. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research*, *58*, 61-68. https://doi.org/10.1016/j.ijer.2012.12.002

Mangesa, R.T., Suhardi, I., & Perenreng, J.M. (2021). An Indonesian case study of computer operating familiarity levels on CBT at vocational high schools. *International Journal of Innovation, Creativity, and Change*, *15*(4). https://www.ijicc.net/images/Vol_15/Iss_4/15417_Mangesa_2021_E1_R.pdf

Martin, R., & Binkley, M. (2009). Gender differences in cognitive tests: A consequence of gender-dependent preferences for specific information presentation formats? In F. Scheuermann & J. Bjórnsson (Eds.). *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing*. Luxembourg: Office for Official Publications of the European Communities. https://publications.jrc.ec.europa.eu/repository/handle/JRC49408

McDonald, A. S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers & Education*, *39*(3), 299-312. https://doi.org/10.1016/S0360-1315(02)00032-5

Mizrachi, D. (2015). Undergraduates' academic reading format preferences and behaviors. *The Journal of Academic Librarianship*, *41*(3), 301-311. http://dx.doi.org/10.1016/j.acalib.2015.03.009

Momeni, A. (2022). Online assessment in times of COVID-19 Lockdown: Iranian EFL teachers' perceptions. *International Journal of Language Testing*, *12*(2), 1-24. https://doi.org/10.22034/ijlt.2022.157122

Neuman, G., & Baydoun, R. (1998). Computerization of paper-and-pencil tests: When are they equivalent? *Applied Psychological Measurement*, *22*(1), 71–83. https://psycnet.apa.org/doi/10.1177/01466216980221006

Ockey, G.J. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing*, *24*(4), 517-537. https://doi.org/10.1177/0265532207080771

O'Malley, K. J., Kirkpatrick, R., Sherwood, W., Burdick, H. J., Hsieh, M.C., & Sanford, E.E. (2005, April). *Comparability of a paper-based and computer-based reading test in early elementary grades.* Paper presented at the AERA Division D Graduate Student Seminar, Montreal, Canada.

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

Paek, P. (2005). Recent trends in comparability studies. *Pearson Educational Measurement Research Report 05*(05). Retrieved from http://images.pearsonassessments.com/images/tmrs/tmrs_rg/TrendsCompStudies.pdf

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. Retrieved on January 18, 2022, from https://eric.ed.gov/?id=ED363272

Plass, J., & Jones, L. (2005). Multimedia learning in second language acquisition. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 467–488). New York: Cambridge University Press.

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), 1–21. Retrieved on May 22, 2021, from https://eric.ed.gov/?id=EJ838613

Pusey, K. (2020). Assessing L2 listening at a Japanese university: Effects of input type and response format. *Language Education & Assessment*, *3*(1), 13-35. https://doi.org/10.29140/lea.v3n1.193

Rafaeli, S., & Tractinsky, N. (2007). Computerized tests and time: measuring, limiting, and providing visual cues for response time in on-line questioning. *Behavior & Information Technology*, *8*(5), 335-351. https://doi.org/10.1080/01449298908914565

Russel, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Educational Policy Analysis Archives*, *5*(3). https://doi.org/10.14507/epaa.v5n3.1997

Russo, A. (2002). Mixing technology and testing. *The School Administrator*, *59*(4), 6–12.

Sangmeister, J. (2017). Commercial competence: Comparing test results of paper-and-pencil versus computer-based assessments. *Empirical Research in Vocational Education and Training*, *9*(3). https://doi.org/10.1186/s40461-017-0047-2

Schulenberg, S. E., & Melton, A. M. A. (2008). The computer aversion, attitudes, and familiarity index (CAAFI): A validity study. *Computers in Human Behavior*, *24*(6), 2620-2638. https://doi.org/10.1016/j.chb.2008.03.002

Smoline, D.V. (2008). Some problems of computer-aided testing and interview-like tests. *Computers & Education*, *51*(2), 745–756. https://doi.org/10.1016/j.compedu.2007.07.008

Sørensen, H., & Andersen, A.M. (2009). How did Danish students solve the PISA CBAS items? Right and wrong answers from a gender perspective. In F. Scheuermann & J. Bjórnsson

(Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing*. Luxembourg: Office for Official Publications of the European Communities.

Triantafillou, E., Georgiadou, E., & Economides, A.A. (2008). The design and evaluation of a computerized adaptive test on mobile devices. *Computers & Education*, *50*(4), 1319–1330. https://doi.org/10.1016/j.compedu.2006.12.005

Vispoel, W. P., Boo, J., & Bleiler, T. (2001). Computerized and paper-and-pencil versions of the Rosenberg self-esteem scale: A comparison of psychometric features and respondent preferences. *Educational Psychological Measurement* *61*(3), 461–74. https://doi.org/10.1177/00131640121971329

Van de Vijver, F.J.R., & Harsveld, M. (1994). The incomplete equivalence of the paper-and-pencil and computerized versions of the General Aptitude Test Battery. *Journal of Applied Psychology 79*(6), 852–59. http://doi.apa.org/journals/apl/79/6/852.pdf

Wagner, E. (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. *Language Learning & Technology*, *11*(1), 67-86. http://dx.doi.org/10125/44089

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. https://www.jstor.org/stable/1435437

Washburn, S., Herman, J., & Stewart, R. (2017). Evaluation of performance and perceptions of electronic vs. multiple-choice paper exams. *Advances in Physiology Education*, *41*(4), 548-555. https://doi.org/10.1152/advan.00138.2016

Wise, S., & Plake, B. (1989). Research on the effects of administering tests via computers. *Educational Measurement: Issues and Practice*, *8*(3), 5–10. https://doi.org/10.1111/j.1745-3992.1989.tb00324.x

Yu, W & Iwashita, N. (2021). Comparison of test performance on paper-based testing (PBT) and computer-based testing (CBT) by English-majored undergraduate students in China. *Language Testing in Asia*, *11*(32), 1-21. https://doi.org/10.1186/s40468-021-00147-0