Chabahar Maritime University
Iranian Journal of English for Academic Purposes ISSN: 2476-3187
IJEAP, 2023, 12(3), 17-35 (Previously Published under the Title: Maritime English Journal)

Computer-Based Testing vs. Paper-Based Testing: Score Equivalence and Testing Administration Mode Preference

[1]Gholamhossein Shahini
[2]Seyyed Morteza Hashemi Toroujeni*

Research Paper IJEAP-2308-1987 DOR: 20.1001.1.24763187.2023.12.3.2.4
Received: 2023-07-21 Accepted: 2023-09-25 Published: 2023-09-27

Abstract: The growing interest in using the technological advantages of Computer-Based Testing (henceforth CBT) over Paper-Based Testing (henceforth PBT) has led to concerns regarding how this transition impact test takers' performanc. The result of such an effect known as testing administration mode effect is the violation of reliability and validity of a test. The equivalency between CBT and PBT is intensively becoming a topic of discussion in educational contexts, especially after the outbreak of Covid-19 pandemic. Investigation of the equivalency of scores procured from two testing modes is required to discover if testing achievement is affected by the transition of testing administration mode. The current research delved into the score equivalence to explore the consistency of test reliability across modes. Moreover, the correlation of testing administration mode preference with testing performance was critically investigated. The findings reported the correspondence of two sets of CBT and PBT scores. Furthermore, sufficient empirical evidence suggested that there was not a statistically significant linear correlation between the testing mode preference and CBT achievement, though most test takers favored CBT. The quantitative research results regarding testing mode preference and CBT attitudes were also underpinned by the results of semi-structured interviews. The current research guides the development of effective strategies to convert PBT to CBT while maintaining the integrity of assessment and ensuring reliability, fairness, and accurate measurement of EFL learners' vocabulary knowledge.

***Keywords**:* Computer-Based Testing, EFL Learners, Paper-Based Testing, Score Equivalence, Testing Mode Preference

Introduction

CBT is increasingly adopted (Chan, Bax, & Weir, 2018) throughout the world (Alkadi & Madini, 2019) due to the advancement of technology and evolving nature of education (Khoshsima & Hashemi Toroujeni, 2017a). Identical CBT and PBT tests may not generate parallel results due to the influence of the "testing administration mode". If similiar scores are obtained from implementing a test in two modes, the test is regarded as consistent and reliable for yielding sustainable results. After the outbreak of COVID-19 and the emerging global pandemic threat throughout the world from the beginning of 2020, CBT was offered to replace PBT by educational institutions regardless of whenever or wherever test takers took their tests for the advantages of CBT such as flexibility and accessibility (Shraim, 2019), automatic scoring, immediate results and instant feedback (Dogan, Kibrislioglu Uysal, Kelecioglu, & Hambleton, 2020), enhanced security, and efficient administration (Khoshsima, Hashemi Toroujeni, Thompson, & Ebrahimi, 2019), and reduced risk of human error (Shute & Rahimi, 2016). According to the Coyne and International Test Commission (2006), the parralelism of two modes' scores should be

[1] Assistant Professor of TEFL, ghshahini@rose.shirazu.ac.ir; Department of Foreign Languages and Linguistics, Shiraz University, Shiraz, Iran.
[2] PhD Student of TEFL (Corrresponding Author), Hashemi.seyyedmorteza@gmail.com; Department of Foreign Languages and Linguistics, Shiraz University, Shiraz, Iran.

verified before the replacement of PBT with CBT because test takers may be at the risk of underperforming in CBT due to the effect of "testing administration mode" (Jabsheh, 2020). Testing administration mode effect signifies that a test developed to measure the same skill in CBT and PBT leads to different scores (Blazer, 2010), and violation of reliability and internal consistency of the test is the subsequent consequence (Hashemi Toroujeni, 2021). Equivalency between CBT and PBT refers to ensuring that two testing modes measure the same skills, knowledge, perfromance, or cognitive abilities or etc. in a comparable manner without violating test's reliability. Equivalency allows for accurate comparison between test takers who choose different modes or encounter other testing conditions (Chen, Cheng, Chang, Zheng, & Huang, 2014). This is important for making fair and informed academic decisions such as college admissions, employment selection, or educational research based on the received scores.

However, there is still not an all-embracing consensus among researchers on an all-inclusive theoretical analysis, examination and explanation for the effect generated by testing mode. Two testing modes encompassing analogous contents should not precipitate significant differences between two sets of scores. Although, Bunderson, Inouye, and Olsen (1989), Retnawati (2015), and Khoshsima and Hashemi Toroujeni (2017b) justified the superiority and popularity of CBT over PBT in their studies, Hardcastle, Hermann-Abell, and DeBoer (2017), and Oz and Ozturan (2018) reported that PBT scores were not substantially different from CBT scores. More equivalent scores result in more reliable tests (Wells & Wollack, 2003). Maintaining the integrity and consistency between two modes is essential to ensure that test takers have an equal opportunity to demonstrate their knowledge, expertise, proficiency, and skills, regardless of the mode in which they are assessed (Wang, Kao, & Chen, 2021). Furthermore, ensuring equivalency in testing conditions underscores educational institutions' dedication to fostering a consistent evaluation process. This commitment extends to the provision of suitable resources and preparation materials, irrespective of the testing mode. Consequently, test takers can engage in preparing effectively and comprehensively for tests, and fostering their confidence and competence, regardless of a specific testing mode. This is fundamentally cricial in ensuring that the assessment accurately measures the intended competencies without any mode-induced bias. Ultimately, adhering to the establishment of equivalency in testing modes and a uniform approach to testing emphasizes institutions' ethical responsibility to offer equitability, integrity, fairness, and excellence in educational evaluations (Gnambs & Lenhard, 2023). Equivalency is especially important in situations where a testing mode needs to be transformed to an alternative version due to the unpredictable circumstances, such as the Covid-19 pandemic have occured recently. If the assessment in two modes lacks equivalency, transitioning between two modes poses difficulties. Establishing equivalency cultivates trust and assurance among testing stakeholders such as test takers, test developers, educators, and policy makers involved in the assessment procedure. This increased confidence leads to more informed decision-making processes and ensures that testing results are reliable measures of test takers' knowledge and skills, rather than being impacted by the mode of testing. Moreover, maintaining equivalency eliminates any potential biases that may arise from differences in testing modes, and promotes fairness and equity in testing process.

In conclusion, the importance of maintaining equivalency between different testing modes, such as CBT and PBT, cannot be overstated. It is essential for building trust among stakeholders, ensuring fair and unbiased testing, and enabling accurate decision-making based on test results. As such, educational institutions and testing organizations should prioritize efforts to establish and maintain equivalency in their testing practices to uphold the integrity and validity of the assessments they administer.

CBT and PBT are deemed reliable and equivalent when they yield similar outcomes through the assessment of comparable content encompassing equivalent knowledge and skills. Studies conducted on the equivalency of CBT and PBT in public schools are inconclusive and limited (Sangmeister, 2017), especially in the EFL domain (Ebrahimi, Hashemi Toroujeni, & Shahbazi 2019) in Iran (Khoshsima and Hashemi Toroujeni, 2017a) and other Asian countries. Consequently, since studies are limited to

evaluating specific measures, recruiting participants of particular contexts, and utilizing different tools to convert PBT into CBT, the related literature may not be generalized to the Iranian public schools' EFL contexts. The current research explored whether PBT and CBT are equivalent in public education in Iran to help accelerate the CBT development.

CBT as the extensively utilized predominant mode of assessment (Jacob, Berger, Hart, & Loeb, 2016) in educational contexts of USA (Chapelle & Voss, 2016), as well as post-industrial and advanced nations (Khoshsima et al., 2019; Schroeders & Wilhelm, 2011), is now undergoing a gradual substitution with PBT in emerging Asian nations like Malaysia, Iran, Jordan, Turkey, etc. (Alakyleh, 2018; Ebrahimi et al., 2019; Retnawati, 2015) due to the inadequate technology infrastructure. Hence, it is crucial to conduct studies that investigate the psychometric equivalence of two modes (Ebrahimi et al., 2019; Schroeders & Wilhelm, 2011). In many Asian nations, the adoption of CBT is not yet widespread (Retnawati, 2015). Most assessment tools used in Asian public education systems, particularly in the fields related to the humanities, continue to rely on conventional PBT methods due to the limited integration of computer-based systems within educational contexts (Komatsu & Rappleye, 2017). In technologically advanced nations like Finland and some leading Asian countries such as Korea, Japan, Hong Kong, and Taiwan, which are globally recognized for their technological advancements (Martin Prosperity Institute, 2015) and high performance in international assessments, the integration of computers within classrooms is remarkably low (Komatsu & Rappleye, 2017). However, it is essential to acknowledge the critical role that computers play in contemporary society and how their profound impact is continuously permeating daily life and contributing to educational success (Alharbi, 2020). Thus, it is reasonable to assess students' academic achievements and document their proficiency using compuetrized methods. This approach aligns with the increasingly digital nature of the world and the educational landscape. (Aydemir, Ozturk, & Horzum, 2013; Singer & Alexander, 2017). However, especially in the context of Iran, despite test developers' acknowledgment regarding the potential benefits of integrating computer technology into test development, there are concerns regarding the integration of CBT in education systems and diverse challenges that CBT could pose to the efficiency of their assessment methodologies.

Literature Review

In retaliation to the interruption in schooling of formal education worldwide due to the increasing spread of Covid-19, Iran's Ministry of Education encountered its century's exceptional learning difficulty was committed to continuing learning through developing the educational network of student. To achieve this goal, the social network and e-learning program known as Shad was publicly released in April 2020 for more than 15 million of Iranian students educated in public schools under the supervision of the Iran Ministry of Education. As a communication and educational software launched by the Ministry of Education of Iran following the outbreak of the coronavirus, more than 80% of Iranian students joined this social network. Iran Ministry of Education announced that students had to follow education through remote learning app called Shad amid coronavirus epidemic, and take their exams through computer or mobile mode until the end of the current educational calendar i.e., June 2021 (The Islamic Republic News Agency, 2020).

In addition to Shad, through which primary and high school students could follow their instructions and exams, virtual classroom software such as BigBlueButton was used to create and host university classrooms in higher education. In line with online education, school, and university students had to take tests in CBT mode. Guidelines for Computer-Based Tests and Interpretations (APA, 1986) expressively asserts that investigation of the parallelism of CBT and PBT scores in terms of psychometrics and statistical strengths of a test (Rausch, Seifried, Wuttke, Kogler, & Brandt, 2016) supported by empirical evidence is a necessity for substituting CBT for PBT (Jeong, 2014). Thereupon,

since preserving the psychometric properties and retaining the integrity of measurement tools is crucial for the success of CBT substitution, the present research aimed to concentrate on whether measurement psychometrics is infringed through transamination of PBT to CBT (TEA, 2008). Translation of PBT into CBT requires CBT mode to be comparable to its PBT counterpart, and scores from the two modes approximate each other. The interchangeability of scores and reliability of two modes of testing is corroborated when similar sets of scores are received from both modes through manipulating strictly coherent, consistent, and systematic experimental evaluation of testing administration mode reconstruction. Conducting comparability investigations helps test designers measure if CBT mode remains reliable and test takers are not disadvantaged by altering mode.

Educationalists, test developers, and instructors must corroborate evaluation of similar knowledge, skill, competency, or proficiency, as well as similar interpretation of scores from CBT and PBT administration modes (Blazer, 2010). Scores must remain unaffected by the mode of testing administration, accurately reflecting test takers' competence and capabilities. Subsequently, opting for the adoption of a particular testing mode becomes more straightforward if the scores from CBT align with those from PBT (Jamieson, 2005). Though Wang, Jiao, Young, Brooks, and Olson (2008) indicated no performance discrepancy between the scores procured from CBT and PBT, Jeong, (2014) and Keng, McClarty, and Davis (2008) reported higher scores on PBT. Additionally, lower scores on PBT were reported by Pommerich (2004). Furthermore, no mode effect was affirmed by Alkadi & Madini (2019), Ben-Yehudah and Eshet-Alkalai (2020); however, Register-Mihalik, Kontos, Guskiewicz, Mihalik, Conder, and Shields (2012) reported testing performance discrepancy between two modes as a result of testing administration mode effect. Since researchers' findings are not conclusive, there is a trend towards conducting comparability studies across different contexts and subject areas (Jabsheh, 2020; Piaw, 2012; Rausch, 2016; Retnawati, 2015), especially in developing countries that have recently commenced replacing PBT with CBT (Ebrahimi, et al., 2019; Khoshsima et al., 2019; Khoshsima and Hashemi Toroujeni, 2017a; Retnawati, 2015).

Investigating whether there exists an association between test takers' inclination for a particular testing mode (CBT or PBT) and their performance in the preferred mode is an area of interest. This inquiry aims to ascertain whether a correlation exists between test takers' testing mode preference and their proficiency in that opted testing mode. By exploring the correlation, researchers seek to understand if test takers' preference for a specific mode impacts their capability to excel in that prefered mode. This understanding helps educationalists make informed decisions regarding the most efficient and fair testing methods. A deep understanding of the interplay between test takers' favored mode and their achievement in testing can profoundly impact how assessments are tailored and administered, ensuring fairness and effectiveness in the evaluation process. A robust correlation between testing mode preference and testing achievement or performance might indicate that incorporating test takers' mode preferences could enhance their performance. Conversely, no correlation between the two variables may signify that factors beyond preference could impact their performance. The association between test takers' testing mode preferences and their corresponding proficiency in their preferred mode carries significant importance. This correlation aids educators and policymakers in making well-informed choices regarding the most effective methods for test administration, with the aim of boosting test takers' performance and mitigating any potential biases introduced by the chosen mode. Some test takers may excel and feel more at ease when assessing in CBT mode for their technological aptitude, while others may gravitate towards the conventional PBT mode.

In addition to assessing the equivalency of CBT and PBT, the present study explored the relationship between test takers' preferences for testing administration mode and their testing performance. The connection between mode preference and testing performance has become increasingly important due to the widespread use of CBT platforms and the expanding integration of educational technology in the field of assessment (Zheng & Bender, 2018). Understanding the potential impact of

administration mode preference on testing outcomes is crucial for educators, policymakers, and test developers because, with the rise of CBT and its integration into educational practices, discerning how learners' inclinations towards CBT or conventional PBT align with their actual performance help educators optimize testing strategies. Subsequently, policymakers and test developers can design assessments that are effective, equitable, and in sync with the preferences and capabilities of modern learners. In essence, acknowledging the impact of administration mode preference on testing outcomes is essential for enhancing educational assessment practices and ensuring fair and meaningful evaluation of learners' knowledge and abilities. Test takers who are more comfortable and familiar with technology might outperform CBT due to their proficiency in navigating digital interfaces and tools (Bennett, Maton, & Kervin, 2008; Hui, Teng, & Guo, 2023). Conversely, test takers who are less comfortable with technology might experience anxiety or confusion that may potentially affect their CBT performance. Testing environment can also impact test takers' performance. Some test takers might prefer the solitude of PBT environment for concentration, while others might outperform in digital environment of CBT with access to search functions and other resources and functionalities. Some test takers might have learning styles that align better with a particular administration mode. Visual learners might prefer CBT with interactive elements, while kinesthetic learners might struggle with screens and prefer hands-on paper-based tests. Some studies reported an association (e.g., Flowers, Do-Hong, Lewis, & Davis, 2011) between testing administration mode preference and testing performance, while others found no significant correlation (e.g., Higgins, Russell, & Hoffmann, 2005; Khoshsima and Hashemi Toroujeni, 2017a; Lightstone and Smith, 2009). As technology continues to shape the ways the learning process is assessed (Yu & Iwashita, 2021), it is crucial to consider how inclination for testing administration mode interacts with testing achievement to minimize bias and enhance the validity of test results. However, the subsequent questions are addressed within the framework of both theoretical and pedagogical perspectives to attain the research aims:

Research Question One: Is there a statistically significant difference between EFL learners' vocabulary achievement in CBT and PBT?

Research Question Two: Is there a statistically significant correlation between test takers' testing administration mode preference and their vocabulary achievement in CBT?

Research Question Three: Do test takers outperform their preferred testing administration mode?

METHODOLOGY

Research Design

A mixed-methods approach synthesizing multiple-choice achievement test, questionnaires, and semi-structured interview within a single-group design was the methodological approach utilized in the current study.

Participants

One-hundred twenty EFL learners of 5 public senior high schools located in Sari, Mazandaran were recruited and assigned to one testing group after the administration of the Oxford Placement Test (OPT) to 189 EFL learners to select homogenous learners, whose scores fell within the range of 120–149 (intermediate level) out of 200 in March, 2023. One-hundred twenty intermediate EFL learners selected for the research objectives included more boys (n=57%) compared to girls (n=43%). Fourteen to seventeen-year-old participants' mean age was 16.5 years with a standard deviation of 1.51.

Instruments

Participants with the same level of EFL proficiency were selected by implementing Allan (2018)'s version of the Oxford Placement Test as a reliable instrument to assess and group research participants based on their level of English language proficiency. The OPT was conducted on a population of 189 Senior High School EFL learners to recruit 120 EFL learners whose scores fell within the range of intermediate language proficiency. In the first testing occasion, the PBT version of the Test 49 (Homes and Buildings/49.2=4 marks, & 49.4=6 marks), Test 60 (Town and Country/60.1=6 marks, 60.2=8marks, 60.3=8 marks, & 60.4=8 marks), Test 63 (Work: Duties, Conditions, and Pay/63.1=4 marks, 63.2=4 marks, 63.3=6 marks, 63.4=6 marks, & 63.5= 10 marks), and test 64 (Jobs/64.1=8 marks, 64.2=8 marks, 64.3=8 marks, & 63.4=6 marks) from Test Your English Vocabulary in Use (Pre-Intermediate and Intermediate) (Redman & Gairns, 2017) was implemented in April 30, 2023 (corresponding to the 10th day of Ordibehesht). Since the tests do not become increasingly difficult in the book, every test is regarded as an independent test and EFL learners are not required to do the tests is a particular order. The tests selected for the current study had a total of 100 marks. Test Your English Vocabulary in Use (Pre-Intermediate and Intermediate) can be used for both pre-intermediate and intermediate EFL learners. The tests are specifically designed for learners at both levels, making them suitable for assessing and reinforcing vocabulary skills at that proficiency levels. However, to ensure the appropriateness of tests for the current research objectives, the current researchers reviewed the whole tests of the book to choose the most appropriate ones regarding their specific instructional goals, curriculum objectives, and lesson contents.

Regarding the research question one, PBT was transformed into CBT using C# programming language within a Windows-based application developed through Microsoft Visual Studio. Microsoft SQL Server was also used for data storage. Test takers had access to the CBT platform by logging in with their unique username and password. Test-takers were provided with a demo that was optional to skip for familiarizing them with the platform and how to take CBT. Test takers' personal information were collected in the initial phase. The test could be initiated by selecting the “Start the Test” button. Each question was presented on a separate sheet, allowing test takers to select a correct option on the screen. The questions were presented to test takers in different orders as multiple sets of question sheets were created. Different sequences of questions were randomly assigned to test-takers automatically. This method ensured enhanced security during testing and minimized the likelihood of cheating. Upon completing the test, results were recorded and displayed by selecting the “Finish” button, which was integrated with Crystal Reports connected to the database. PBT papers were manually evaluated by the researchers. If a test taker marked multiple options in PBT, no score was assigned to the question.

In the second testing occasion, CBT was implemented in May 21, 2023 (corresponding to the 30th day of Ordibehesht), after three-week interval between two testing occasions. Participants were instructed to carefully read each question presented individually on the screen and then select their answers from the provided options for that question. They could determine the most suitable option as the answer by clicking on the blank space beside the options. If needed, a thoughtful review of answers could be efficiently conducted by ticking the multiple-choice boxes. Reviewing responded items required navigating through multiple pages as each question was displayed individually on a separate page. Test takers were required to answer one hundred multiple-choice test questions within eighty minutes.

To address the second research question, the question i.e., “Would you prefer taking the test on paper –no difference– on computer” was presented to test takers at the foot of exam paper and exam screen to scrutinize the interdependence between two variables of testing administration mode preference and testing achievement.

Furthermore, to address the research question three, the mean scores across different preference groups obtained from PBT (pre-CBT) and CBT (post-CBT) sessions were used to delve into whether there was a correlation between the test takers' preferred testing mode and their actual performance. For instance, if

the CBT mean score was higher for a preference group that favored CBT (On Computer Prefernce Group) compared to the PPBT mean score, it would suggest a potential correlation between preference and performance. Conversely, if the mean scores did not differ across preference groups, it might indicate that test takers did not necessarily outperform in their preferred mode. This data is crucial for addressing Research Question 3, which explores the relationship between test takers' performance and their mode preference, providing valuable insights into how preferences influence actual test outcomes.

Moreover, to delve into the test takers' perspectives regarding testing modes, a researcher-constructed questionnaire (Test Takers' Attitudes) was administered to the test takers subsequent to their exposure to CBT. This thoughtfully designed instrument, detailed in Table 5, aimed to comprehensively evaluate the attributes of both PBT and CBT. These questions were consistently presented to all test takers, ensuring a standardized approach for direct comparison of responses. The questions were structured with fixed options (on paper, no difference, on computer), ensuring a consistent and systematic approach to collecting data, allowing the test takers to select their responses from the provided predefined choices. Hence, in addition to exploring the third research question, TTA (Test Takers' Attitudes) questionnaire was administered to the test takers to gauge their perspectives on particular aspects of CBT and PBT. The aspects were derived from an extensive review of the related literature and researchers' insights. The questionnaire included an evaluation of ten distinct features of the tests.

Subsequently, twenty volunteer test takers participated in in-depth semi-structured interviews after collecting their informed consent forms. With a one-week interval after the implementation of CBT, semi-structured interviews were conducted to delve into not only test takers' preferences regarding testing administration modes but also particular features of CBT and PBT. "Can you elaborate on your preference for a specific testing mode, whether it's CBT or PBT?", and "what were the key reasons behind your preference?" were the semi-structured interview questions aimed to delve deeper into the factors that shape test takers' preferences and attitudes towards testing modes, offering insights into their experiences and rationale behind their choices.

Procedure

To explore the equivalency of CBT and PBT, the intermediate EFL learners were administered the PBT in April, 2023. After a three-week interval, they were given the CBT version in May. To investigate the correlation between testing administration mode preference and test takers' performance, the simple testing mode preference question was asked at the end of both PBT and CBT exams. At the termination of CBT (second session) in May, 2023, test takers responded to the TTA questionnaire concerning their preference for testing mode. Table 5 presented below offers a comprehensive assessment of ten aspects related to testing administration modes, aiming to shed light on the preferences and experiences of test takers. The table's ten questions probed different dimensions of the test-taking experience of participants, and their preferences were assessed through percentages. The first question delved into the navigation dimension, examining which testing mode offered smoother navigation for questions and items. Subsequent questions investigated the readability and comprehensibility of questions and items, the level of fatigue induced during test completion, the ease of recording answers, and the straightforwardness of reviewing and altering answers. Additionally, the table explored the test-takers' comfort levels during testing administration and their perceptions of score consistency and enjoyment in different testing modes. Lastly, it assessed the precision of vocabulary knowledge measurement. The provided percentages represented the preferences of respondents for each question across three categories: On Paper, No Difference, and On Computer. These responses provided valuable insights into whether test takers exhibited preferences for one mode over another in terms of the aspects evaluated.

With a one-week interval after the implementation of CBT, semi-structured interviews were conducted. The interview analysis commenced with transcribing the recorded dialogues. Two TEFL

professors validated the accuracy of the transcriptions. Then, a thematic analysis was performed to draw meaningful concepts, which were then categorized under common themes (Seidman, 1998). To ensure the credibility of the themes, a member-check approach was employed to validate the themes developed based on the researchers' subjective experiences. Twelve interviewees were invited to confirm whether the extracted pieces accurately reflected their interview responses. Last but not least, the content validity, cohesion, coherence, and accuracy were assessed by two TEFL experts. Through a triangulation of data and integrating a multifaceted methodology (both questionnaires and interviews), a thorough understanding of test takers' perspectives on testing modes was obtained. This multifaceted methodology reinforced the strength and reliability of the conclusions.

Results and Discussion

Using SPSS 22, the overall reliability coefficients for testing the internal consistency of two testing modes were calculated using the scores received from 120 participants. The estimated internal consistency demonstrated high Cronbach's alpha coefficients (PBT/α=.83 & CBT/α=.85) and verified reliability. Additionally, the normality distribution of data was statistically evaluated using Shapiro-Wilks and Kolmogorov-Smirnov. For PBT and CBT, p-values of .821, and .873 were obtained, respectively, demonstrating that research data followed a normal distribution. The results of Levene's Test ($F(1,119) = 8.3$, $p = 0.0$), with an alpha level at 0.05 ($p(0.758) > \alpha$ .05), were obtained to examine the homogeneity of variances. According to the results, both assumptions of data normal distribution and homogeneity of variances as prerequisites of conducting parametric statistical tests were met. The highest mean score in PBT (M=46.85, SD=1.43), surpassing CBT (M=46.13, SD=1.8) by 0.72 points, demonstrated that test takers outperformed PBT. Moreover, the larger standard deviation of PBT scores indicated a wider distribution of scores compared to CBT. Conversely, the narrower spread of CBT scores suggested that the scores were concentrated around the CBT mean.

Moreover, the standard error showed that the calculated values for both modes were approximately true values. The *standard* error indicated a precision in the calculated values. It suggested that the sample mean was likely close to the actual population mean. A smaller standard error signifies that the estimate is reliable and has a higher probability of accurately representing the population parameter. The standard error of the mean (SEM) of 1.07 in Table 1 signifies that there is a level of uncertainty or variability associated with the estimated mean difference between the PBT and CBT. The SEM of 1.07 indicated that the sample mean difference of 46.70 is likely to vary by approximately 1.07 units. A smaller standard error generally suggests that the sample mean is a more accurate representation of the true population mean. In this context, the SEM of 1.07 implied that the estimated difference in scores between PBT and CBT could be reasonably precise, with a range of approximately ±1.07 units. However, it's important to note that a smaller standard error does not necessarily imply a smaller margin of error, but rather a more precise estimation of the mean difference. Then, the low SEM was an indicator of the tests' reliability.

To find a statistically significant difference between the CBT and PBT mean scores received from implementation of two testing modes, the results of Paired-Sample T-test (Table 1) demonstrated no statistically significant difference at the .05 significance level, indicating the probability of risk in estimating the difference between the two mean scores. As a result, the significance value of .565 at a significance level of $P<0.05$ with 119 degrees of freedom (N-1) indicated that there was no statistically significant difference between the two means (Sig=.565, $P>0.05$).

Table 1

Paired T-Test Results for PBT and CBT

	Paired Differences					t	D.F.	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
PBTCBT	46.70	1.34	1.07	-3.04097	5.44097	.579	119	.565

In analyzing the paired differences between PBT and CBT for the EFL learners' vocabulary achievement, the mean difference was found to be 46.70, suggesting that, on average, participants scored slightly higher in PBT compared to CBT. However, this difference was within a relatively narrow range, with a standard deviation of 1.34, indicating that the variation in the differences was not significant. The standard error of the mean difference was calculated at 1.07, underlining the precision of the estimated difference. Additionally, the 95% confidence interval for the difference (-3.04097 to 5.44097) implied that the true difference in vocabulary achievement between the two modes could reasonably fall within this range. This provided evidence against a statistically significant difference between the means of PBT and CBT. The t-value (0.579) yielded a non-significant p-value of 0.565 (two-tailed) with 119 degrees of freedom. This confirmed the absence of statistical significance and supported the conclusion that there was no significant difference between the vocabulary achievement scores obtained through PBT and CBT.

These results *suggested* that, within the context of this study, the mode of testing—whether PBT or CBT—does not significantly impact EFL learners' vocabulary achievement. The minor difference in mean scores indicated that both testing modes were comparable in evaluating vocabulary knowledge in this specific EFL context. The absence of a significant difference in scores between these modes suggested that both testing approaches measured vocabulary knowledge similarly. The implications of the findings imply that educators and test developers can choose either mode, CBT or PBT, for vocabulary assessment. This flexibility allows for adaptability in testing methods, considering factors such as available resources, technological infrastructure, or test takers' preferences. However, it's crucial to acknowledge that the current study's results may be context-specific and may not necessarily be generalizable to all educational settings. The comparability of CBT and PBT could vary based on multiple factors, including the nature of the content being tested, the level of learners, and the specific skills being assessed. Future research should continue exploring this comparability in diverse contexts to build a comprehensive understanding of the relationship between testing modes and language assessment outcomes.

The correlation of test takers' responses to the testing administration mode questionnaire implemented in two testing occasions including PBT (pre-CBT preference) and CBT (post-CBT preference) with their mean score on CBT was tested. The results from Pearson's product-moment correlation analysis, carried out using SPSS to assess the expected correlation between pre-CBT mode preference ($r=0.013$, $n=118$, $p<0.821$) and post-CBT mode preference ($r=0.019$, $n=118$, $p<0.632$), with the CBT performance of the test-takers, indicated a weak positive correlation (Table 2) that was not statistically significant.

Table 2

Pearson Correlation Coefficients for Pre-CBT and Post-CBT Mode Preference

	Pearson Correlations	Pre-CBT Mode Preference	Post-CBT Mode Preference
CBT Performance	Pearson Correlation	.013	.019
	Sig. (2-tailed)	.821	.632

N	120	120

The Pearson correlation coefficients indicated weak positive correlations between both pre-CBT and post-CBT mode preferences and CBT performance. These findings suggested that there was only a marginal relationship, if any, between test-takers' preferences for a specific mode of testing and their actual performance in the CBT environment.

The lack of statistically significant correlation in this context revealed that a preference for a specific testing mode, whether CBT or PBT, does not reliably predict test takers' vocabulary achievement in the CBT environment. The common belief may be that EFL learners would outperform in a mode they prefer or are familiar with. This study rejects this assumption by suggesting that factors other than mode preference may significantly impact vocabulary achievement in CBT. Since understanding what influences vocabulary achievement in CBT is critical, this finding encourages a more critical evaluation of the factors influencing CBT performance, potentially shifting the focus towards computer literacy, proficiency in navigating the CBT interface and understanding its tools and functionalities, vocabulary knowledge, test-taking strategies, question types, or test design, and etc. Testing is a complex process influenced by multiple or multidimensional factors. According to the results, it is not merely about how test takers prefer to take a test, but also about their actual knowledge, comprehension, and adaptability to the testing environment. Then, this study emphasizes that reducing this complexity to a mode preference might oversimplify the assessment process.

From an educational perspective, these findings underscore the importance of preparing students for diverse testing environments. Rather than focusing solely on a single mode, educational institutions should equip learners with the skills and adaptability needed to excel in both CBT and conventional PBT. The findings ensure that EFL learners are not disadvantaged by their mode preference and can perform optimally in various assessment contexts. Moreover, in addressing the third research question, multiple comparisons utilizing descriptive statistics of different groups of testing administration mode preference were employed to probe deeper into the correlation between mode preference and testing performance.

Table 3

PBT Mean Score of Testing Administration Mode Preference Groups (pre-CBT)

Pre-CBT Mode Preference	N	PBT Mean Score	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
On Paper PG.	85	38.54	2.43	1.32	26.5481	43.0469	32.00	40.00
No Difference PG.	15	43	1.56	.265	37.6735	44.3427	38.00	46.00
On Computer PG.	20	51	.01	.01	50.0000	51.0000	50.00	51.00
Total	120	44.18	1.33	0.531	38.0738	46.1298	40.00	45.00

PG. Preference Group

Based on the outcomes of comparing mean scores among different preference groups derived from pre-CBT assessments associated with the respondents' preferences for testing mode administration, it was observed that the test takers who expressed a stronger preference for CBT achieved the highest mean PBT score (pre-CBT/On-Computer PG.'s PBT/M=51, (SD=.01)), implying that those favoring CBT outperformed test takers who exhibited a preference for PBT (pre-CBT/On-Paper PG.'s PBT/M=38.54, (SD= 2.43)) in PBT session. Accordingly, the greater PBT mean score of the On-Computer preference

group compared to the On-Paper preference group indicated that testing administration mode preference did not necessarily lead to better performance on the preferred testing mode (Table 3). On the contrary, test takers advocating for PBT performed better in CBT (post-CBT/On-Paper PG.'s CBT/M=39.64, (SD=1.59)) in comparison to their performance in the PBT session (pre-CBT/On-Paper PG.'s PBT/M=38.54, (SD= 2.43)) (Table 4). However, proponents of CBT displaying a preference for this mode, although exhibited superior performance in CBT compared to On Paper Prefernce Group with PBT administration mode preference (On-Paper) in the CBT session, did not surpass in its favored mode ((pre-CBT/On-Computer PBT/M=51, (SD=.01)) vs post-CBT/On-Computer CBT/M=42.06, (SD=.02)). However, although proponents of CBT who showed their preference for CBT outperformed CBT compared to other testing administration mode preference groups (On-Paper and No-Difference) in post-CBT survey, they failed to outperform their favored testing mode (pre-CBT/On-Computer PG.'s PBT/M=51, (SD=.01) vs post-CBT/On-Computer PG.'s CBT/M=42.06, (SD=.02)). In contrast, supporters of PBT in PBT session (pre-CBT/On-Paper PG.'s PBT/M=38.54, (SD=2.43)) (Table 3) outperform their CBT (post-CBT/On-Paper PG.'s CBT/M=39.64, (SD=1.59)) (Table 4) with a more excellent mean score. Similarly, the performance of test takers who showed no strong preference for a particular testing mode (No-Difference preference group) excelled in CBT (post-CBT/No-Difference PG.'s CBT/M=47.35, (SD=2.03)) compared to their PBT performance (pre-CBT/No-Difference PG.'s PBT/M=43, (SD=1.56)), as well as the performance of other preference groups in the CBT session (Table 4). However, the analysis of mode preferences across different groups revealed no significant correlation between the preferred testing mode and test takers' performance in the test.

Table 4

CBT Mean Score of Testing Administration Mode Preference Groups (post-CBT)

Post-CBT Mode Preference	N	CBT Mean Score	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
On Paper PG.	85	39.64	1.59	2.01	36.31569	43.6327	12.00	54.00
No Difference. PG	15	47.35	2.03	1.51	38.3157	52.0897	32.00	48.00
On Computer PG.	20	42.06	.02	.02	41.0000	42.0000	41.00	42.00
Total	120	43.01	1.21	1.18	38.5437	45.9074	28.00	48.00

PG. Preference Group

The findings indicated that test takers' tendency towards a specific testing mode did not necessarily lead to a superior performance in the preferred mode. This suggested that a variety of different factors beyond testing mode preference, such as study habits, skill levels, exam preparation, familiarity with the mode, overall cognitive abilities, and etc. may significantly influence testing performance. While a preference for a particular testing mode may influence test takers' comfort level during the test, it does not consistently impact their actual testing performance. However, the current results do not substantiate superior performance based on testing mode preferences.

Lastly, the study investigated the test takers' perceptions of ten aspects related to CBT and PBT, including (a) ease of navigation, (b) clarity of text (readability), (c) level of fatigue, (d) ease of recording answers, (e) reviewing and (f) modifying answers, (g) comfort in the testing environment, (h) confidence in score consistency, (i) enjoyment of the testing experience, and (j) accuracy and precision in measuring. Table 5 illustrates the frequency and percentage of test takers' preferences for these features.

Table 5

Test Takers' Attitudes towards PBT and CBT

No.	questions	On Paper%		No Difference%		On Computer%	
		F.	P.	F.	P.	F.	P.
1	Which test provided smoother navigation for questions and items?	25	20.83	32	26.66	63	52.5
2	In which test were questions and items more legible and easier to understand?	18	15	30	25	72	60
3	Which test induced less fatigue during completion?	18	15	28	23.33	74	61.66
4	In which test was recording answers a more straightforward task?	21	17.5	25	20.83	74	61.66
5	In which test was reviewing given answers a more straightforward process?	31	25.83	14	11.66	75	62.5
6	Which test facilitated more straightforward alteration of answers?	10	8.33	18	15	92	76.66
7	Which test was characterized by a higher level of comfort during administration?	35	29.16	15	12.5	70	58.33
8	In which test is it more likely that your score would remain consistent if you were to retake it?	35	29.16	20	16.66	65	54.16
9	Which test did you find enjoyable to participate in?	25	20.83	20	16.66	78	65
10	Which test provided a precise measurement of your vocabulary knowledge?	9	7.5	11	9.16	100	83.33

Analyzing the test takers' responses, as shown in Table 5, revealed that 20% of them found navigating the PBT environment easier, whereas 52% of the total 120 test takers favored the ease of navigation in CBT. In terms of legibility, 60% of the test takers found CBT more efficient in terms of item comprehension, processing, differentiation, and interpretation, in contrast to the 15% who endorsed PBT for its readability. These test takers expressed confidence that their preferred mode facilitated convenient reading of questions and options. Notably, 61% of the respondents praised the revitalizing testing environment and digital interface of CBT, while only 15% of the test takers found favoured less fatiguing level of PBT testing mode. Moreover, 61% favored CBT due to its capability for recording and submitting responses conveniently, while less than 18% regarded PBT as efficient enough for easy recording of responses.

Additionally, a majority of 62% of respondents found CBT easier to review their responses, while 25% expressed a preference for PBT in terms of reviewability. Notably, the results indicated that 76% of participants perceived CBT as more conducive to modifying their answers, while less than 10% reported a higher comfort level with PBT for altering responses. Furthermore, 58% of tes takers found the ergonomic design of the CBT environment more comfortable, while less than 30% indicated a stronger inclination towards using PBT for the purpose of comfortability. Besides, 54% of the participants expressed confidence that their scores would remain consistent upon retaking CBT, in contrast to a minority of 29% who leaned towards choosing PBT for achieving consistent results in a subsequent administration. Regarding enjoyment, 20% favored PBT, while a majority of 65% found CBT to be more enjoyable. Notably, 83% perceived CBT as a more reliable tool for assessing their vocabulary knowledge, while only 7% expressed confidence in PBT for achieving comparable accuracy in vocabulary assessment.

The twenty interviewees were asked to provide detailed insights into their attitudes towards CBT and PBT. Among them, a majority of 80% expressed a strong preference for CBT, while a minority of

20% leaned towards PBT. Notably, the preferences did not align with the test takers' actual performance, a conclusion drawn from the quantitative investigations. In line with the quantitative results, the qualitative analysis demonstrated no correspondence and consistency between test takers' testing mode preference and their CBT performance. The results of the quantitative analysis demonstrated that the test takers who were inclined towards PBT (Table 3) outperformed CBT (Table 4). Conversely, supporters of CBT (table 3) excelled in their PBT. Based on the qualitative results, the highest rate of interviewees preferred CBT. CBT advocators expressed several benefits to clarify their choice for CBT.

Interviewees' justifications corresponded to their responses to the testing administration mode preference question and the TTA questionnaire exploring their perspectives on the features of PBT and CBT. All interviewees supporting CBT cited advantages such as "the ease of reading items", "the ease of selecting and modifying answers", and "immediate access to scoring reports". Furthermore, 78%, 60%, and 57% of CBT supporters highlighted the features of "enhanced security", "faster decision-making enabled by immediate scoring", and "the efficiency in terms of less time and effort to take CBT", respectively, for CBT. Despite the majority of interviewees expressed a preference for CBT, some of them still favored conventional testing. For instance, 100% of PBT supporters emphasized the advantages like "easy navigation", "familiarity with the testing format", "ease of circling questions and answers for later review", and "the absence of additional task demands". In the context of CBT, additional task demands were cited by PBT supporters to refer to extra actions, and requirements such as navigating through the CBT interface, understanding the complex instructions to take the CBT such as the optional demo presented to them in the initial phase of CBT, managing the digital testing environment, and using specific software features that they needed to perform beyond the fundamental tasks and above simply answering the test questions. However, 85% of PBT supporters raised concerns about the time-consuming nature of answer review in CBT, primarily due to the presentation of a single question on the screen and the need to navigate through multiple pages for reviewing specific questions.

Multiple studies utilizing different samples and methodologies have been conducted to explore the effect of CBT or PBT administration mode on EFL learners' performance. Hashemi Toroujeni (2021) revealed that there was no statistically significant difference between EFL learners' testing performance on CBT and PBT. While some researchers suggested that CBT leaded to slightly higher scores due to its interactive nature (e.g., Alkadi & Madini, 2019; Aydemir et al., 2013; Singer & Alexander, 2017; Wang et al., 2021), others, in agreement with this study, indicated no significant difference in CBT and PBT outcomes (e.g., Chan et al., 2018; Hashemi Toroujeni, 2021; Jamieson, 2005; Jeong, 2014; Yu & Iwashita, 2021). The findings of Chen et al. (2014), and Bennett et al. (2008), who demonstrated lower scores in CBT and confirmed incomparability of CBT and PBT are against the findings of the current study. These conflicting findings emphasize the necessity of conducting further research. Several moderator variables, such as individual learning styles, test anxiety, familiarity with technology, etc. can influence testing performance, regardless of testing mode. In comparability studies in which a significant difference is found between CBT and PBT, the effect of such moderator variables should be explored.

The results of the current study demonstrated that the correlation between test takers' testing administration mode preference and testing performance is not always straightforward, and test takers do not surpass their preferred testing mode. Test-takers who are more comfortable with technology and have positive attitudes towards CBT may show more preference for CBT. They may be more accustomed to the digital interfaces, which could positively impact their navigation and interaction in CBT environments. Conversly, test takers accustomed to conventional PBT assessments may struggle with the digital version of tests due to the factors such as unfamiliarity with technology, or the need to adapt to new navigation methods. Based on the results of this study and Lightstone and Smith (2009), there is not always a strong correspondence between test takers' preferred testing mode and their actual testing performance. The current study revealed that test takers may prefer a testing mode due to their familiarity or literacy, but they do not outperform in their preferred testing mode.

Instead of solely focusing on specific mode preferences, test takers could be provided with a variety of testing experiences that expose them to both CBT and PBT. This approach can help test takers become adaptable to different testing conditions, preparing them for a diverse range of assessments they may encounter in their academic contexts. Furthermore, future research could delve deeper into the factors contributing to testing performance. By understanding the interplay between cognitive factors, preparation strategies, and individual preferences, educators can refine their teaching methods and assessment strategies to better support all learners' learning outcomes.

Conclusion and Implications

In conjunction with the progressive acceptance of computer technology in Iran, especially after the Covid-19 outbreak and temporary closure of schools for face-to-face education as a result of the emerging global pandemic threat, the current study delved into the equivalency of CBT and PBT in Public Senior High Schools, and the correlation of testing administration mode preference with testing performance. Statistical analysis revealed no significant difference in test scores across the modes, and the testing administration mode preference survey indicated that test takers' performance was not generally correlated with their testing performance because they did not outperform their preferred testing mode. Furthermore, the attitude survey demonstrated that most test takers held positive attitudes toward CBT. Thematic analysis of the transcribed data received from interview sessions showed that test takers preferred CBT for its advantages over PBT, such as "enhanced security", "faster decision making as a result of immediate scoring", less time and effort to take CBT", "easy to read items", "easy to choose answers", "easy to change answers", and "immediate scoring reports" which increase their testing efficiency. Test takers who preferred PBT cited "easy to navigate", "more familiarity with the testing format", being accustomed to circle the questions and answers for later review", and "no need to extra task demand" as the benefits of PBT.

CBT and PBT can effectively evaluate test-takers knowledge and skills on the same content. The questions, prompts, and tasks in both modes can cover the same topics and levels of difficulty. However, CBT and PBT can be considered equivalent and reliable regarding similar scores received from assessing the same content knowledge and skills through using similar question types. Nevertheless, the equivalency of CBT and PBT is a multidimensional issue and the choice between them depends on factors such as test-takers' technological proficiency or digital literacy, desired level of immediate feedback, and the logistical feasibility of each mode that should be considered by educational institutions. Both modes have their strengths and limitations, and the choice between them should be based on careful consideration of the factors. Educational institutions must address the challenges of each testing mode when transitioning from PBT to CBT. PBT might be logistically challenging when dealing with a great number of test takers. It provides more familiarity with testing format and limited technical literacy, but lacks the convenience of swift scoring and feedback. While CBT offers great benefits, such as enhanced accessibility, immediate feedback, more effortless scalability, personalized and tailored testing experiences through adaptive algorithms that tailor questions difficulty hinged on test takers' current abilities, it is not without its technical and anxiety-related challenges. Educational institutions must explore these factors carefully when choosing a testing mode and test takers may also consider their own preferences and technological proficiency to choose one. As the educational landscape evolves, the future of testing will likely to involve hybrid models that provide the best of both modes.

While test takers may express preferences for either PBT or CBT based on their perceived strengths in particular testing mode, it is important to note that their personal preferences do not significantly impact their actual testing performance. The research findings, which indicated no statistically significant difference between the performance of EFL learners in CBT and PBT, carry significant implications for language education. Educators can use the insights provided by the current

research to use technological advancements, and integrate CBT into their teaching and testing methodologies. Furthermore, educational institutions and testing organizations are required to maintain the equivalency of both testing modes. This is especially important for preserving the validity and reliability of test scores, ensuring that the integrity of assessments remains intact.

Acknowledgement

Special thanks to the Sari Department of Education for their cooperation in conducting the current research.

Funding Details

No funders were involved in study design, analyses, manuscript preparation, or decision to submit for publication.

Conflicts of Interest/Competing Interests

The authors have no competing interests to declare that are relevant to the content of this article. The authors have no relevant financial or non-financial interests to disclose.

Experimentation Consent of Participants

To comply strictly with research ethics and to give enough awareness of the nature of the study to the students, all of them were given a consent form to sign. They were told to declare their agreement to participate in the study.

References

Alakyleh, A. S. (2018). Evaluating the Comparability of (PPT) and (CBT) by Implementing the Compulsory Islamic Culture Course Test in Jordan University. *International Journal of Assessment Tools in Education*, 5(1), 176-186. https://doi.org/10.21449/ijate.370494.

Alharbi, L. (2020). The Effectiveness of Using Interactive Technology and Video Games on Developing English as a Foreign Language among Saudi Students in the Qassim Region. *TESOL International Journal*, 15 (5): 6-30.

Alkadi, S. Z., & Madini, A. A. (2019). EFL learners' Lexico-grammatical competence in paper-based vs. computer-based genre writing. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3431758.

Allan, J. (2018). Oxford Placement Test. Oxford University Press.

American Psychological Association (APA), (1986). *Guidelines for Computer-Based Tests and Interpretations.* Washington, DC: Author.

Aydemir Z., Ozturk E., & Horzum, MB. (2013). The effect of reading from screen on the 5th grade elementary students' level of reading comprehension on informative and narrative type of texts. *Educational Sciences: Theory and Practice,* 13(4):2272–2276. DOI: 10.12738/estp.2013.4.1294

Bennett, S., Maton, K. & Kervin, L. (2008). The "Digital Natives" Debate: A Critical Review of the Evidence. *British Journal of Educational Tech*nology, 39(5), 775-786. https://doi.org/10.1111/j.1467-8535.2007.00793.x.

Ben-Yehudah, G., & Eshet-Alkalai, Y. (2020). Print versus digital reading comprehension tests: does the congruency of study and test medium matter?. *British Journal of Educational Technology*, 0(0). https://doi.org/10.1111/bjet.13014.

Blazer, C. (2010). *Computer-Based Assessments* (Vol. 0918). INFORMATION CAPSULE Research Services. Retrieved from: https://files.eric.ed.gov/fulltext/ED544707.pdf.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), *Educational Measurement* (pp. 367–407). Washington, DC: American Council on Education. https://doi.org/10.1002/j.2330-8516.1988.tb00291.x.

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

Chapelle, C. A., & Voss, E. (2016). 20 years of technology and language assessment in language learning & technology. *Language Learning & Technology, 20*(2), 116–128. Retreived from: https://www.learntechlib.org/p/176101/.

Chen, G., Cheng, W., Chang, T. W., Zheng, X., & Huang, R. (2014). A comparison of reading comprehension across paper, computer screens, and tablets: Does tablet familiarity matter? *Journal of Computers in Education, 1,* 213–225. https://doi.org/10.1007/s40692-014-0012-z.

Dogan, N., Kibrislıoglu Uysal, N., Kelecioglu, H., & Hambleton, R. K. (2020). An overview of e-assessment. *Hacettepe University Journal of Education*, 35(Special Issue), 1-5. DOI: 10.16986/HUJE.2020063669

Ebrahimi, M.R, Hashemi Toroujeni, S.M., & Shahbazi, V. (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 (iJET)*, *14*(07). https://doi.org/10.3991/ijet.v14i07.10175.

Flowers, C., Do-Hong, K., Lewis, P., & Davis, V. C. (2011). A comparison of computer-based testing and pencil-and-paper testing for students with a read- aloud accommodation. *Journal of Special Education Technology*, 26(1), 1-12. https://doi.org/10.1177/016264341102600102.

Gnambs, T., & Lenhard, W. (2023). Remote Testing of Reading Comprehension in 8-Year-Old Children: Mode and Setting Effects. Assessment, 0(0). DOI: 10.1177/10731911231159369

Hardcastle, J., Hermann-Abell, C. & DeBoer, G. (2017). *Comparing Student Performance on Paper-and-Pencil and Computer-Based-Test*. Retrieved from: https://files.eric.ed.gov/fulltext/ED574099.pdf.

Hashemi Toroujeni, S.M. (2021). Computerized testing in reading comprehension skill: investigating score interchangeability, item review, age and gender stereotypes, ICT literacy and computer attitudes. *Educ Inf Technol* 27, 1771–1810 (2022). https://doi.org/10.1007/s10639-021-10584-2.

Higgins, J., Russell, M., & Hoffmann, T. (2005). Examining the effect of computer-based passage presentation on reading test performance. *Journal of Technology, Learning, and Assessment*, 3(4). Retrieved from: https://ejournals.bc.edu/index.php/jtla/article/view/1657.

Hui L., Teng L.S., & Guo F. (2023). Modeling the relationship between digital nativity and Smartphone usage in learning English as a foreign language contexts. *Front. Psychol*. 13:1053339. https://doi.org/10.3389/fpsyg.2022.1053339.

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

Jabsheh, A. H. M. (2020). The Usability Outlook of Computer-Based Exams as A means of Assessment and Examination: A case study of Palestine Technical University. *International Journal of Linguistics, Literature and Translation (IJLLT)*, 3(3). DOI: 10.32996/ijllt.2020.3.3.16.

Jacob, B., Berger, D., Hart, K. C., & Loeb, S. (2016). *Can technology help promote equality of educational opportunities?* In K. Alexander & S. Morgan (Eds.), The Coleman report and educational inequality fifty years later (pp. 242-271). New York, NY: Russell Sage Foundation.

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. *Behaviour & 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. (2017a). Computer-Based Testing: Score Equivalence and Testing Administration Mode Preference in a Comparative Evaluation Study. *International Journal of Emerging technologies in Learning*, 12 (10), pp. 35-55. https://doi.org/10.3991/ijet.v12i10.6875.

Khoshsima, H. & Hashemi Toroujeni, S.M. (2017b). Transitioning to an Alternative Assessment: Computer-Based Testing and Key Factors related to Testing Mode. *European Journal of English Language Teaching*, 2 (1), pp. 54-74. http://dx.doi.org/10.46827/ejel.v0i0.499.

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 (TEwT)*, 19(1), 86-101. https://files.eric.ed.gov/fulltext/EJ1204641.pdf.

Komatsu, H. & Rappleye, J. (2017). Did the shift to computer-based testing in PISA 2015 affect reading scores? A View from East Asia. *Compare: A Journal of Comparative and International Education*, 47:4, 616-623. https://doi.org/10.1080/03057925.2017.1309864.

Lightstone, K., & Smith, S. M. (2009). Student Choice between Computer and Traditional Paper-and-Pencil University Tests: What Predicts Preference and Performance? *International Journal of Technologies in Higher Education*, 6(1), 30-45. Retrieved from: https://www.erudit.org/fr/revues/ritpu/2009-v6-n1-ritpu3631/039179ar.pdf.

Martin Prosperity Institute. (2015). *"Creativity and Prosperity: The Global Creativity Index 2015."* Toronto: Martin Prosperity Institute. Retreieved from: http://boletines.prisadigital.com/Global-Creativity-Index-2015.pdf.

Oz, H., & Ozturan, 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. Retrieved from: https://www.jlls.org/index.php/jlls/article/view/878.

Piaw, C. (2012). Replacing paper-based testing with computer-based testing in assessment: Are we doing wrong? *Procedia - Social and Behavioral Sciences, 64*, 655–664. https://doi.org/10.1016/j.sbspro.2012.11.077.

Pommerich M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *The Journal of Technology, Learning, and Assessment*, 2(6) (2004). Retrieved from: https://ejournals.bc.edu/index.php/jtla/article/view/1666.

Rausch, A., Seifried, J., Wuttke, E., Kogler, K., & Brandt, S. (2016). Reliability and validity of a computer-based assessment of cognitive and non-cognitive facets of problem-solving competence in the business domain. *Empirical Research in Vocational Education and Training*. 8(9). https://doi.org/10.1186/s40461-016-0035-y.

Redman, S., & Gairns, R. (2017). *Test Your English Vocabulary in Use Pre-Intermediate and Intermediate*. Cambridge University Press

Register-Mihalik, J. K., Kontos, D. L., Guskiewicz, K. M., Mihalik, J. P., Conder, B., & Shields, E. W. (2012). Age-related differences and reliability on a computerized and a paper-pencil neurocognitive assessment battery. *Journal of Athletic Training, 47*(3), 297–305. doi: 10.4085/1062-6050-47.3.13.

Retnawati, H. (2015). The Comparison of Accuracy Scores on the Paper and Pencil Testing vs. Computer-Based Testing. *TOJET: Turkish Online Journal of Educational Technology*, 14(4). Retrieved from: http://www.tojet.net/articles/v14i4/14413.pdf.

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.

Schroeders, U., & Wilhelm, O. (2011). Equivalence of Reading and Listening Comprehension across Test Media. *Educational and Psychological Measurement*, 74(5). https://doi.org/10.1177/0013164410391468.

Seidman, I. (1998). *Interviewing as Qualitative Research: A Guide for Researchers in Education and the Social Science*s (2nd Ed.). New York: Teachers College Press.

Shraim, K. (2019). Online Examination Practices in Higher Education Institutions: Learners' Perspectives. *Turkish Online Journal of Distance Education*, 20(4), 185-196. DOI: 10.17718/tojde.640588

Singer, L. M. & Alexander, P. A. (2017a). Reading on Paper and Digitally: What the Past Decades of Empirical Research Reveal. *Review of Educational Research*, *87*(6), 1007–1041. https://doi.org/10.3102/0034654317722961.

Singer LM, & Alexander PA. (2017b). Reading across mediums: Effects of reading digital and print texts on comprehension and calibration. *The Journal of Experimental Education*, 85(1):155–172. DOI: 10.1080/00220973.2016.1143794

Shute, V.J & Rahimi, S.A. (2016). Review of Computer-Based Assessment for Learning in Elementary and Secondary Education. *Journal of Computer Assisted Learning*. 33 (1), 1-19. https://doi.org/10.1111/jcal.12172.

Texas Education Agency. (2008). *A review of literature on the comparability of scores obtained from examinees on computer-based and paper-based tests*. Retrieved from: https://tea.texas.gov/system/files/2008-LiteratureReviewComparabilityReport.pdf.

The Islamic Republic News Agency. (2020, May*). Eight-four Percent of Iranian Students are Members of Shad*. Retrieved from: https://www.irna.ir/news/83813020/.

Wang, T.-H.; Kao, C.-H.; Chen, H.-C. (2021). Factors Associated with the Equivalence of the Scores of Computer-Based Test and Paper-and-Pencil Test: Presentation Type, Item Difficulty and Administration Order. *Sustainability*, 13, 9548. https://doi.org/10.3390/su13179548.

Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 assessment: A meta-analysis of testing mode effects. *Educational and Psychological Measurement*, 68, 5-24. https://doi.org/10.1177/0013164406288166.

Wells, S.C. & Wollack, J.A. (2003). *An Instructor's Guide to Understanding Test Reliability*. Testing & Evaluation Services publication, University of Wisconsin. Retrieved from: https://testing.wisc.edu/Reliability.pdf.

Yu, W., & Iwashita, N. (2021). Comparision 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.

Zheng, M., & Bender, D. (2018): Evaluating outcomes of computerbased classroom testing: Student acceptance and impact on learning and exam performance, *Medical Teacher*, 41(1):75-82. DOI: 10.1080/0142159X.2018.1441984