Language Testing in Asia

RESEARCH Open Access

Comparison of test performance on paper-based testing (PBT) and computer-based testing (CBT) by English-majored undergraduate students in China

Wenjing Yu* and Noriko Iwashita

* Correspondence: wenjing.yu@uq.net.au
School of Languages and Cultures, Faculty of Humanities and Social Sciences, The University of Queensland, Brisbane, Qld 4072, Australia

Abstract

Computer-based testing (CBT), which refers to delivering assessments with computers, has been widely used in large English proficiency tests worldwide. Despite an increasing CBT in China, limited research is available concerning whether CBT can be used for the Test for English Majors-Band 4 (TEM 4). The current study investigated whether testing mode impacted TEM 4 score and factors (i.e., computer familiarity level and attitude towards CBT) that might correlate with performance on CBT of TEM 4. Overall 92 Chinese undergraduate students were randomly assigned to one of the groups, i.e., CBT or paper-based testing (PBT), and took the test. A mixed method was employed, including (1) quantitative and qualitative analysis of test performance in two modes, as well as CBT group participants' computer familiarity and attitudes towards the mode; and (2) thematic analysis of semi-structured interviews. The results revealed that (1) test scores in CBT and PBT were comparable; (2) two items in the computer familiarity questionnaire, i.e., comfort level of reading articles on the computer and forgetting time when using computers, positively correlated with CBT scores; and (3) participants' attitude towards CBT did not impact test performance.

Keywords: Computer-based testing, Paper-based testing, Test performance, Score equivalence, Computer familiarity, Test takers' attitude, Test mode preference

Introduction

Background

The history of exams and tests in China can be traced back to the imperial period some 2000 years ago (202 BC–220 AD) (Bai, 2020). The early 1980s witnessed the principles and practice of modern tests in China, following reform and opening-up policy and the National College Entrance Examination (NCEE) scheme, commonly referred to as Gaokao (Li, 2019). Examination in China is generally hailed as an effective and unbiased approach to selecting talents, enhancing teaching quality, and even fighting against outright corruption and nepotism in allocating limited societal resources

(Yan & Fan, 2020; Zhang, 2021). As China has seen growing participation in the international community, English is increasingly significant for English as a Foreign Language (EFL) learners at all stages of education (Jin, 2010). The Chinese Ministry of Education (MOE) has organized many different Fan examinations at both provincial and national levels, including the College English Test (CET) and Test for English Majors (TEM) (Li, 2019).

Test for English Majors-Band 4

Test for English Majors-Band 4 (TEM 4) is a national English proficiency test specifically intended for English-majored undergraduate students in tertiary education in China (Fan et al., 2020), and the annual test is used to evaluate students' English proficiency and achievement in listening, reading, and writing (Jin & Yan, 2017).

Over the past few years, TEM 4 has gained plenty of social respect as an important marker of English achievement, with its score serving as an essential criterion for English majors to acquire a bachelor's degree or even to graduate from university (e.g., Jin & Yan, 2017; Qian & Cumming, 2017). In addition, passing the TEM 4 test is one of the requirements for obtaining a residence permit in some metropolises such as Shanghai, Shenzhen, and Beijing in China (Yu & Jin, 2014).

Listening, reading, writing skills are assessed in TEM 4 via various online and offline modes ranging from traditional single-choice and multiple-choice questions to more integrative formats, including dictation, note-taking, and gap-filling (Jin & Fan, 2011). In the listening and reading sections, test takers' linguistic knowledge such as vocabulary and syntactic knowledge and foundational cognitive skills such as working memory and attention are assessed (e.g., Florit et al., 2013; Tompkins et al., 2013). Furthermore, the evaluation of higher order cognitive skills is also involved: test takers need to evaluate initial and local propositions through comprehension, then cross-check propositions and fill in missing information using inferencing skill (the ability to identify missing information in the text based on prior knowledge) and theory of mind (the ability to infer others' mental states and predict behavior) (Kim, 2016; Kim & Phillips, 2014). The writing section is designed to evaluate their ability of transcription, which refers to the process of converting sounds into written symbols, including spelling and handwriting skills, and text generation, which consists of generating and organizing ideas (Juel et al., 1986), as well as oral language representation (Berninger et al., 2002; Kim et al., 2011). The test has been administered via paper and pencil since its official launch in 1992(Jin & Fan, 2011) (Additional file 1).

Computer-based test

With a spurt of progress in technological tools, computer-assisted learning has emerged as an important aspect of language teaching, and computer-based test (CBT) has been gradually introduced as a result (Kim et al., 2018). Computer technology is being used "in designing, developing, and delivering test content as well as scoring and reporting examinee test performance" (Sawaki, 2012, p. 426). Srivastava and Gray (2012) considered CBT as the representative of the tests in the 21st century, explaining that an increasing number of large-scale tests have been administered in this mode. CB language tests in their early forms mainly focused on short response items or multiple-choice questions. However, open-ended questions like essay writing have become more common in recent years due to technological advancements and shifts to include more authentic tasks (e.g., Choi et al., 2003; Tajuddin & Mohamad, 2019; Wang & Kolen, 2001).

The claimed benefits of CBT have been a major driving force behind its implementation. Firstly, CBT can provide useful information about the test takers' instructional needs. For example, the test takers' process of answering questions helps identify mistakes and misunderstandings and provides adequate instructions in their test preparations. (Jiao & Wang, 2010; Kunnan, 2015; Patel & Laher, 2011). Further, instructors can monitor the distribution of time they spend answering questions, which offers them insights into students' instructional needs (Anakwe, 2008; Tajuddin & Mohamad, 2019). CBT also provides a unique test-taking experience. Compared to a traditional paper-based testing (PBT), CBT can more easily include tests with embedded multimedia, resulting in simulated experiences that students may respond to and interact with (Prisacari & Danielson, 2017). Embedded assistive technology options available in CBT can provide specific functions such as closed captioning and text-to-speech, allowing test takers with physical disabilities to complete the test independently (Boo & Vispoel, 2012; Kolen, 1999).

Score equivalence of two modes

In the past few decades, a growing number of studies have focused on the score equivalence of CBT and PBT, as well as test takers' feedback on the two modes (e.g., AlKadi & Madini, 2019; Banerjee, 2003; Chapelle, 2001; Goldberg & Pedulla, 2002; Patel & Laher, 2011; Prisacari & Danielson, 2017; Randall et al., 2012; Wang et al., 2007). Replacing PBT with CBT requires interchangeable test score data based on a comparative study of the two modes. If the scores are not identical, or at least close, the score difference is considered as the "test mode effect" (Fulcher, 1999; Goldberg & Pedulla, 2002). International Guidelines on CBT stated that when an assessment is implemented in two modes and two sets of similar scores are obtained, the scores are considered equivalent and reliable (International Test Commission, 2006). The equivalent test scores established for the two modes demonstrate valid and reliable CBT. According to the classical True-Score Theory, the same test conducted via CBT and PBT should result in equivalent or identical test scores (Khoshsima et al., 2017). Therefore, as the first step to develop and administer a CBT of TEM 4 at universities in China, a comparative study on the effect of the two test modes on test takes' performance should be conducted to explore whether the two sets of scores are comparable and consequently valid and reliable.

Previous studies on language tests have shown various results when it comes to the effect of the two test modes on test takers' performance. Some studies found that students who took PBT outperformed those who took CBT or vice versa (e.g., AlKadi & Madini, 2019; Coniam, 2006). For example, according to Choi et al. (2003), a total of 365 students from five universities in Korea took both PBT and CBT, but half of them took PBT first, then CBT, and the other half took both in a reversed order. The results revealed that students scored higher in the listening task in CBT than in PBT. The presence of visual information on CBT, according to the researchers, may explain such significant score differences in the listening test. Similarly, in a collaborative writing test administered in PBT and CBT, AlKadi and Madini (2019) compared sentence-level errors and lexicon-grammatical ability. The results showed that 73 intermediate-level female university students in Saudi Arabia who took the PBT made more sentence-level errors than those who took the CBT. The researchers further explained that test takers can make fewer errors in CBT than in PBT thanks to the functions of the computer such as auto-correction and smart prediction.

However, the equivalence of PBT and CBT test performance was found in other studies (e.g., Chapelle, 2007; Jamieson, 2005; Kim et al., 2018). For example, Mohammadi & Barzgaran, 2012 investigated the comparability of a writing test administered in PBT and CET, respectively in Iran. The data on the test performance of 80 undergraduate EFL learners on the writing section of a Cambridge Preliminary English Test (PET) revealed no statistically significant difference between their PBT and CBT writing scores. Similarly, Öz and Özturan (2018) investigated the score equivalence of an achievement by 97 EFL teachers in Turkey and found comparable scores between the two modes. Finally, according to Yao (CitationID2019), 289 secondary school EFL students from Macau took the two modes of the Test of Academic English (TAE) at a local university, with no statistically different scores between CBT and PBT.

Influencing factor of CBT performance

Computer familiarity level has been identified as an important influencing factor of test performance (Chan et al., 2018). As CBT requires the skills of online editing and keyboarding, participants' test results may be affected by their familiarity with using a computer (Yeom & Jun, 2020). Hence, CBT may become a disadvantage for learners who were unfamiliar with CB testing interfaces or had less access to computers.

For instance, in assessing the comparability of Graduate Record Examinations (GRE) administered in PBT and CBT respectively, Goldberg and Pedulla (2002) found that (1) PBT test takers outperformed CBT test takers; and (2) those who were more familiar with computers received higher scores than those who were not. This study investigated computer familiarity via a questionnaire survey of test takers' familiarity with various computer functions. Likewise, Odo (2012) studied whether computer familiarity of EFL secondary school students in Canada affected their performance on a reading test. A computer familiarity questionnaire was performed for 120 individuals, followed by an online multiple-choice reading test, and the results revealed that the familiarity level of participants has a minor but significant impact on the test scores.

However, some studies reported no links between computer familiarity and test scores. For example, Hosseini et al. (2013) studied the correlation between computer familiarity and test scores on an institutional English reading test attended by 162 Iranian EFL learners from four different educational institutes. Participants in the CBT and PBT were given two comparable reading assessments, followed by a questionnaire on computer familiarity to understand their levels of computer familiarity. The findings showed no significant influence of their computer experience and skills on the reading comprehension test results. More recently, Khoshsima et al. (2019) evaluated 58 adult EFL learners' CBT and PBT results in Iran to see whether test takers' computer familiarity will influence their CBT scores. The results showed no statistically significant correlation between computer familiarity and CBT scores.

In addition to computer familiarity level, the test takers' preference for delivery mode is also a critical factor in comparability studies (e.g., Hosseini, 2017; Khoshsima et al., 2017; Uke, 2017). Studies on CBT test takers' attitudes, like those on score equivalence and computer familiarity, yielded conflicting results. For example, Higgins et al. (2005) investigated the attitudes of 219 fourth-grade primary school students towards CBT reading comprehension tests. Even though participants preferred CBT over PBT, the test results for the two modes were identical. In the same vein, Khoshsima et al. (2017) and Ebrahimi et al. (2019) also found that CBT test takers with positive attitudes towards the mode failed to earn higher test scores.

However, Flowers et al. (2011) studied the preference of students with disabilities for PBT and CBT, as well as their test results, and ultimately, the opposite results were obtained. Data from a large-scale testing program was used to investigate possible differences in academic performance between CBT and PBT for third to eleventh-grade EFL learners with a read-aloud accommodation in math, reading, and science. 77% of the participants believed that they performed better in CBT than in PBT. However, PBT test takers, despite their higher preference for CBT, outscored CBT test takers in the reading test, indicating a negative correlation between their attitudes towards CBT and the test scores.

Present study

Previous studies on the score equivalence of the two modes, computer familiarity, and attitude towards CBT revealed varied results, which may be caused by specific factors such as study context differences, contextual differences (e.g., test types and task difficulties), and test taker group differences (e.g., age and gender), hence more researches in specific contexts are required (Anakwe, 2008; Kunnan & Carr, 2017).

The possibility of shifting traditional PBT to CBT requires to be studied due to the vast and growing number of Chinese EFL learners. Although most studies on Chinese CBT discussed the possibility, few studies reported empirical findings (Yu & Zhang, 2017). The impact of the test mode on secondary school students' scores was investigated in the most recent study regarding the score equivalence of the two modes (Yao, 2019). Little has been done from the perspective of Chinese undergraduate students. Based on the review of the previous studies, the difference in test taker population is a key element affecting test performance, thus to fill the gap identified in the literature review above, the following three questions are addressed:

- Do students' performance on reading, listening, and writing sections of TEM 4 differ between the two test modes?
- To what extent does CBT test takers' computer familiarity affect their test scores?
- What are the attitudes of CBT test takers towards the test mode they received and why?

Methodology

Participants

This study included 92 second-year English-majored undergraduate students aged 19–21 (82 females and 14 males) from a large Chinese university, where they learned English language, literature, and linguistics, and had studied English for some 11–13 years at the time of data collection. The unequal gender distribution of the sample is consistent with that in many English departments in Chinese universities (Jin, 2010; Yu & Jin, 2014). All test takers shared the same native language background, Mandarin Chinese, and scored between 6.0 and 6.5 in IELTS. They value their performance in TEM 4 as passing the test is a requirement to acquire their bachelor's degree (Yan & Fan, 2020).

Since socioeconomic status (SES) is often associated with access to learning resources such as computers (Walpole, 2003), teachers asked their students about their SES before participating in this study. To identify the family financial status of students, Table 1 was designed based on the data from the National Bureau of Statistics of China (NBSC) (http://www.stats.gov.cn/english/PressRelease/202007/t20200716_1776358.html). According to the NBSC, the per capita disposal income in China reached 15666 yuan in the first half-year of 2020, which was used to dope out the remaining intervals of the scale (see Table 1). Participants were divided into two social and economic groups, with 72% of them in the lower-middle class (*N* = 66) and 28% in the upper-middle class (*N* = 26). They were therefore assumed to have full access to computers.

Participants were randomly divided into two groups: group 1 (*N* = 46) and group 2 (*N* = 46) to participate in CBT and PBT of TEM 4, respectively. As shown in Table 2, since the data was not normally distributed, a Mann-Whitney *U* test was performed to check the equivalence of the English language proficiency between the two groups. The non-significant difference indicates the comparable English proficiency of the two groups.

Instruments

TEM 4 tests in two modes, computer familiarity, and attitude questionnaires, as well as three semi-structured interview questions, were prepared for data collection. Computer

Table 1 The category of total monthly family incomes (unit: RMB yuan)

	Low income	Lower-middle income	Middle income	Upper-middle income	High income
Family income range	< 7834	7834–15666	15667–23499	23500–31322	> 31322

Table 2 IELTS scores of group 1 and group 2 (distribution and Mann-Whitney test)

	N	*M*	*SD*	Kurtosis	Skewness	*df*	*P*
Group 1	46	6.24	0.25	− 2.09	.25	45	.16
Group 2	46	6.22	0.25	− 2.02	.25		

familiarity and attitude surveys were performed for group 1 participants who took the CBT. Semi-structured interviews were also conducted for group 1 participants only.

TEM 4

The testing materials in this study were extracted from the original official TEM 4 tests issued in 2018 and 2019 and was downloaded from an online educational website called Baidu Wenku (https://wenku.baidu.com/view/367b02e23a3567ec102de2bd960590c69ec3d8a1.html). In this test, the reading section was extracted from the official 2018 TEM 4 test, including five articles with 15 multiple-choice questions, and the listening section was extracted from the listening comprehension sections of both the official 2018 and 2019 TEM 4 tests, including 30 multiple-choice questions. The writing section was extracted from the official 2019 TEM 4 test, and participants were required to write approximately 300 words on a given topic. The test tasks in the writing, reading, and listening sections in the two test modes were identical.

Computer familiarity questionnaire

The online computer familiarity survey written in English was adapted from a study by Weir et al. (2007), with some modifications to keep up with advances in computer technology and to reflect the current research context (see Appendix A), including 11 Likert scale questions which may take 10–15 min to complete. Items 1, 2a, and 2b inquired about participants' computer usage frequency; items 3 to 5 investigated their comfort levels when reading and typing on computers; items 6 to 9 examined their thoughts on taking TEM 4 on a computer, and the final item asked participants to rate their computer familiarity level. The computer familiarity questionnaire was provided to group 1 participants one day before the test.

A pilot test was conducted to examine the validity of these questions. A sample of 28 group 1 participants (9 males and 19 females) took the test on WeChat, a popular Chinese mobile texting and voice messaging communication application. During the survey, any issues with participants' understanding of the terms, the length of the survey, etc., were recorded and then adjusted accordingly in the final version of the questionnaire.

The perceived task values scale consisting of 10 questionnaire items was subjected to reliability analysis, and Cronbach's alpha (.82) revealed that the questionnaire was reliable. All items should be kept and removing them will result in a reduction in the alpha.

Attitude questionnaire

Following the survey used by Escudier et al. (2011), the online questionnaire written in English was composed of two sections (see Appendix B). The six statements (items 1 to 6) in section A about attitudes towards CBT of TEM 4 were expected to be rated by a 5-point Likert scale, ranging from strongly agree (1) to strongly disagree (5). In section B, a multiple-choice question (item 7) was set to explore test takers' overall attitude towards CBT, followed by an attitude survey 5 min after the test, which required to be completed in less than 15 min.

Like the computer familiarity survey, a pilot test of the attitude questionnaire was conducted for group 1 participants, with a Cronbach's alpha of .84 indicating its reliability. Most items should be kept and removing them will result in a reduction in the alpha, except for item 8 which will increase the alpha to .86.

Data collection

Test administration and questionnaire survey

The TEM 4 test was conducted at the university where test takers were studying at the time of data collection. All participants were given clear instructions before taking the test and their responses were kept confidential. The test in both modes was conducted by the two instructors teaching College English courses at the university responsible for playing the listening texts in the listening section. Both CBT and PBT test takers were given 90 min to complete the test.

One day before the test, group 1 participants completed the computer familiarity questionnaire. In CBT (group 1), participants were required to type the essays in the writing section on the computers, with automatic word counts and no proofreading function (e.g., spell-checkers and grammar-checkers), and their answer sheets were submitted automatically when the time was up. Test takers were not allowed to review the completed sections after set-up (e.g., 30 min for reading). After a 5-min break, they were asked to participate in the attitude survey, and all responses were saved automatically after submission.

Participants in PBT (group 2) were given three answer sheets of listening, reading, and writing sections. Like the CBT, the test monitor was responsible for playing the listening texts, and collecting according to answer sheets after the set-up time of each section (e.g., 30 min for reading), so test takers could not review the completed sections. The answer sheet of the writing section was collected when the 90 min were up.

Semi-structured interview

The data collected from the online survey may not reveal the unforeseen reasons why participants prefer CBT of TEM 4 or not. Therefore, three days after the CBT, three open-ended questions (see Appendix C) were given to 10 randomly selected group 1 test takers as a semi-structured interview. Question 1 examined their perception of CBT before taking the test, while question 2 inquired about their attitude towards CBT after the test. In question 3, participants needed to explain why their attitude remained changed or unchanged. The interview was conducted via Zoom (a software program offering audio/video chatting service) in English and was expected to end in about 15 min. All interviews were audio-recorded, with the recordings being transcribed by the researcher.

Data analysis

Scoring test performance

The test scores of the listening and reading sections of CBT of TEM 4 was dichotomously marked (i.e., 0 for the wrong answer and 1 for right answer), which were recorded automatically via the Sojump platform, a professional online questionnaire, voting and, evaluation platform in China and downloaded as Excel files. The answer sheets for the reading and listening sections in PBT were manually marked by the two monitors, who were also responsible for entering the results of the multiple-choice questions in PBT to obtain an Excel file.

For the PBT writing test, to avoid rater bias due to test takers' handwriting, their essays were first typed into the computer and checked by the researcher, and the scores of both the typed and handwritten texts were combined as one data set. Two monitors independently rated the texts on the computer using a 20-point rubric, which offers descriptions for five scores, 2, 6, 10, 14, and 18, in terms of discourse cohesion, language quality, and content relevance (see Appendix D). A third rating was conducted by the researcher for essays with a difference of three points, with the mean of the two adjacent scores as the final result.

Inter-rater reliability

The scores of the PBT writing section by the two raters were strongly correlated (see Table 3), with the same trend found in the CBT writing scores, indicating that the scores of both writing tests were reliable.

Examination of research questions

For RQ1, a multivariate analysis of variance (MANOVA) was performed with SPSS 26.0 to determine whether there were differences in the test performance on the listening, reading, and writing sections (dependent variables) between groups 1 and 2. A significance level of .05 ($p \leq .05$) was set to explore whether there was any statistically significant difference in test takers' performance between the two testing modes.

For RQ2, the means of 10 items in the computer familiarity survey were calculated via SPSS 26.0, followed by a Pearson correlation analysis to examine the relationship between the various aspects of group 1 participants' computer familiarity and their CBT performance.

For RQ3, the descriptive statistics of the items in the attitude survey were presented and compared. As for the interview data, an inductive approach was applied for thematic analysis as outlined by Braun and Clarke (2006), which consists of six steps. Step 1 is to be familiar with the entire data set, and this study used a voice-recognition software named Dragon voice recognition (https://www.nuance.com/dragon.html) to quickly obtain the transcripts, and then checked the transcripts against original audio recordings for accuracy. In step 2 (generating initial codes), the potential data items of interest, questions, connections between data items, and other preliminary ideas were recorded. After recording how codes developed from ideas, the same codes were applied to the entire data set by labeling data extracts with relevant codes and making note of any potential patterns or connections between items that may inform subsequent theme development (Braun & Clarke, 2006). The third step involves an

Table 3 Inter-rater reliability

	PBT writing R2	CBT writing R2
PBT writing R1	.92[a]	0
CBT writing R1	0	.93[a]

[a] Correlation is significant at the 0.01 level (2-tailed)

examination of the coded and collated data extracts to identify potential themes of broader significance (Braun & Clarke, 2006). In this step, the thematic maps were used to visualize cross-connections between concepts and among main themes and subthemes. In step 4 (reviewing themes), the entire data set was firstly re-examined to ensure common and coherent data within each theme, and that data between themes was distinct enough to merit separation (Braun & Clarke, 2006). Additional data under the themes that had been newly created or modified were then re-coded and the thematic map was accordingly revised. After the refinement of the thematic map in step 5, the overlap between themes was identified and the scope of what each theme entails was delimited (Braun & Clarke, 2006). In step 6, the final analysis and description of findings were completed.

Results

Comparison of test performance (RQ1)

As Table 4 shows, the test performance of CBT test takers (group 1) and PBT test takers (group 2) are not significantly different across the sections and in the total score. The test score of group 1 participants is slightly higher than that of group 2 except for the listening section.

Before running the MANOVA analysis, the figures of Kurtosis and Skewness showed that the assumption of univariate normality was satisfied (see Table 8 in Appendix E), and the results of Levene's test of homogeneity of variances indicated that the assumption of homogeneity of variances was also satisfied (see Table 9 in Appendix E). MANOVA analysis showed no statistically significant difference in performance between CBT and PBT, F (3, 98) = .66, p = .6; Wilks' Λ = .88 (see Table 10 in Appendix E), indicating the equivalence of test performance between the two test modes.

Computer familiarity level (RQ2)

The means and standard deviations of test taker responses to 11 items in the computer familiarity questionnaire are shown in Table 5. The means ranged from 3.35 to 4.22 on a 5-point Likert scale, indicating that group 1 participants appear to be familiar and comfortable with using computers.

Pearson correlation analysis was conducted to investigate the possible association between test performance and varied aspects of computer familiarity, with the results summarized in Table 6. The analysis revealed that only two items, the comfort level of

Table 4 Overall descriptive statistics by sections in CBT and PBT

Section	Full mark	Group	*M*	*SD*
Listening	30	CBT	22.7	3.46
		PBT	22.8	3.77
Reading	15	CBT	11.3	2.36
		PBT	10.9	2.26
Writing	20	CBT	13.2	3.05
		PBT	13.1	3.01
Total score	65	CBT	47.2	8.12
		PBT	46.9	8.45

Table 5 Descriptive statistics of the computer familiarity questionnaire

Question		*M*	*SD*
1	How often do you use a computer?	4.22	1.01
2a	How often do you use a computer for entertainment?	3.35	1.14
2b	How often do you use a computer for studying?	3.98	.68
3	How comfortable are you when using a computer in general?	4.13	.78
4	How do you feel about using the keyboard (typing)?	4.07	.83
5	How comfortable are you when using a computer to read an article?	4.15	.94
6	Taking exams on the computer is an inevitable trend.	3.67	.79
7	My computer skill is enough to take CB language tests.	3.76	.74
8	I think taking exams on a computer is interesting.	3.72	0.78
9	When I am working with the computer, I always forget about the time.	4.07	1.02
10	Compared to other students, I rate my familiarity with a computer at...	4.22	1.01

reading articles on the computer (item 5) and forgetting the time when using a computer (item 9), had a minor but significant positive correlation with test takers' scores in CBT.

Attitude towards CBT (RQ3)

Attitude survey

As explained earlier, Section A (items 1 to 6) of the attitude survey presented five aspects that might affect students' perception of taking TEM 4 on a computer, among which items 1, 2, and 3 asked about their perception of listening, reading, and writing sections in CBT, items 4 and 5 examined their comfort level of taking CBT. Item 7 (*What is your overall attitude towards the computer-based version of TEM 4: A. Positive; B. Negative; C. Neutral*) in Section B of the survey examined their overall attitude towards CBT.

As shown in Table 7, students believed that compared to PBT, CBT made it easier to (1) browse reading passages (item 1); (2) listen to and select the answer at the same time (item 2); and (3) edit and reorganize their written texts (item 3). However, for the comfort level of taking CBT, participants found other test takers' typing sounds

Table 6 Pearson correlation analysis of CB score on questionnaire items

Question		*r*	Sig. (2-tailed)
1	How often do you use a computer?	.16	.29
2a	How often do you use a computer for entertainment?	.04	.77
2b	How often do you use a computer for studying?	.04	.81
3	How comfortable are you when using a computer in general?	.27	.08
4	How do you feel about using the keyboard (typing)?	.06	.68
5	How comfortable are you when using a computer to read an article?	.33	.03
6	Taking exams on the computer is an inevitable trend.	.001	.1
7	My computer skill is enough to take CB language tests.	.13	.4
8	I think taking exams on a computer is interesting.	.06	.73
9	When I am working with the computer, I always forget about the time.	.31	.03
10	Compared to other students, I rate familiarity with a computer at...	.05	.75

Table 7 Descriptive statistics of items A to F (*N* = 46)

Section A		*M*	*SD*
1	I find the reading passages are easy to browse in CBT of TEM 4.	3.07	1.03
2	In CBT of TEM 4, I can edit my essay more easily than with pencil and paper.	3.24	1.08
3	I find it easy to listen to and type at the same time in the listening section in CBT of TEM 4.	3.43	1.09
4	I am not distracted by the typing sounds made by other test-takers.	2.67	1.03
5	Taking CBT does not make me more anxious.	2.43	1.03
6	Compared with PBT, I feel more relaxed while taking CBT.	2.48	1.09

distracting (item 4) and believed that this new test mode made them nervous and anxious (items 5 and 6).

For item 7, out of 46 test takers, 24 enjoyed taking CBT, while 12 did not, with 10 holding a neutral opinion. Therefore, half of the participants had a positive attitude towards taking tests on a computer.

Semi-structured interview

The interview involves three questions, with questions 1 and 2 asking about participants' attitudes towards CBT before and after the test, respectively, and question 3 examining why their mode preference remained changed or unchanged. The results showed that 6 participants (60%) did not prefer CBT and 4 (40%) held a positive opinion before the test; while 6 participants (60%) preferred CBT after the test. Three themes were repeated in their responses, including efficiency, test experience, and test environment.

As efficiency was frequently mentioned as an important criterion for the degree of satisfaction of test takers for CBT, a theme was developed to address the connection between these two concepts. Participants preferring CBT explained why they were more efficient in CBT in terms of listening and writing sections. In the listening section, as reported in the results of the attitude survey above, participants found it easy to click the test items and choose or change answers on a computer screen during a listening task. Excerpt 1 below further illustrates this point.

> Excerpt 1: The listening section is difficult as people in the conversations all speak very fast, and I can hardly follow. In the past, I could not write down my answer while listening carefully. However, in CBT, I just needed to click my answer, which saved my time and was very effective. (Interviewee C)

As for the writing section, participants preferring CBT explained that they could structure and edit their essays more easily in CBT than in PBT. They mentioned that in PBT, they were required to use ink pens as handwritten scripts in ink could generate visually high quality of writing sufficient for onscreen marking. Therefore, they were unwilling to amend what they had written on paper due to the concern of "the neatness of their texts" (Interviewee A). Interviewees further explained that they had more time to "reorganize their sentences" (Interviewee F) and "produce longer essays in CBT" (Interviewee G), as it was more convenient for them to type with a keyboard or revise the existing content than writing by hand. Excerpt 2 shows the opinion of Interviewee F on the effectiveness of writing articles in CBT.

> Excerpt 2: I think CBT improved my efficiency which is the key to success. I can delete phrases or type new words until I feel that the sentence is right. I can't edit my essay freely in PBT as it will make my paper messy. Additionally, I can complete my essay quicker in CBT than in PBT due to my fast typing, leaving more time for me to reorganize my essay. (Interviewee F)

Participants not preferring CBT explained that it was inefficient to read and understand passages on a computer as they were used to reading online for pleasure rather than analyzing or extracting information from the text. Interviewee G found it difficult to perform the reading section of TEM 4 on a computer as she could not concentrate on decoding the meaning of the passages.

> Excerpt 3: CBT lowered my efficiency in performing the reading test. I didn't know how to complete a reading test online. My eyes turned away from the screen uncontrollably and I can't focus on the reading tasks. I guess this is because I am used to reading online just for entertainment. (Interviewee G)

Test experience was another theme that emerged in test takers' responses. Some expressing a positive attitude towards CB test experience explained that taking CBT was "innovative and enjoyable" (Interviewee C) and regarded CBT as "a more comfortable test experience than PBT" (Interviewee H). They also mentioned the accuracy of CBT, as it eliminated the human error in the scoring of multiple-choice questions and improved the quality and reliability of the test. However, some thought CBT increased their test anxiety: Interviewee A stated that "I never took this mode of test before, and I am stressed... ". Interviewee I, who has only taken the traditional PBT before, stated some reasons for regarding taking CBT as an unpleasant experience as shown below.

> Excerpt 4: I don't feel nervous if it is a test mode I have experienced before. CBT is brand-new, so I naturally become anxious. I felt awful during the exam because I spent too much time on the questions, I knew that it was because of anxiety, and eventually, I ran out of time. (Interviewee I)

The final theme was the test environment. Participants preferring CBT found that it offered a less fatiguing and more enjoyable environment due to certain screen elements such as colors, graphics, and text together. Yet some participants thought that the CB test environment was disappointing as they were heavily distracted by the noises made by other test takers. Interviewee J stated as below when he was asked about the reason for disliking the CBT test environment.

> Excerpt 5: I don't like the testing environment. We should take the test in a larger classroom. I hate the typing noise made by others. I also think the keyboards in our university are too old, and we need to press the keys very hard. (Interviewee J)

In summary, test takers preferring CBT (N = 6) believed that the mode enhanced their efficiency in listening and writing sections, offered them an innovative test experience, and created a comfortable test environment. However, those not preferring CBT

(N = 4) stated that CBT lowered their speed at analyzing texts, increased their test anxiety, and distracted them due to others' typing sounds.

Discussion

Comparison of scores by section (RQ1)

The MANOVA analysis revealed no statistically significant difference in scores in reading, listening, writing sections, and the whole test between the two modes, and many studies supported the findings of this study (e.g., Chan et al., 2018; Jamieson, 2005). The comparable test results between CBT and PBT indicate the possibility of CBT serving as a feasible solution to respond to the trend of employing computers in language testing and meeting students' and institutions' needs. One possible explanation for the score equivalence of the two modes is that participants are rather familiar with using computers thanks to the computer class each week and new word learning via online resources in daily tutorials. Recent years have witnessed a growing introduction of computers in academic contexts, allowing an increase in participants' familiarity with and opportunity for using computers in testing, classroom activities, and everyday life, which might have impacted their test performance and result in similar scores (Chan et al., 2018). For further studies, it would be better if speaking could also be taken into consideration. Although speaking plays a significant role in language testing, only a few studies provided evidence of score equivalence between two modes, face-to-face and videoconferencing, on the testing of oral skills.

Test takers' computer familiarity level (RQ2)

The computer familiarity questionnaire indicated that participants were generally familiar with using computers, which might be attributed to their SESs as explained in the methodology section. According to Battle and Lewis (2002), the SES of learners served as a fundamental factor that may contribute to English language learning outcomes. In this case, students in the middle or high social and economic classes may perform equally well in CBT and PBT due to their abundant learning resources which are necessary to cope with the new computerized testing format. In contrast, those with lower SESs usually deal with a lack of resources and might have limited access to computers, resulting in lower scores in CBT. In future research, learners with varied educational or social backgrounds (e.g., major, university, and region) should be recruited, promoting the findings to be generalized to the language testing system of higher education in China.

Furthermore, the positive correlation between test takers' computer familiarity and scores was not found in previous studies (e.g., Hosseini et al., 2013; Khoshsima et al., 2019), indicating that computer literacy is an evolving notion given the rapid advancement of diverse digital platforms and technologies, thus persisting attention should be given by the test designers. Multimodal transmission of language test materials via digital technology such as audio, visual, and animations, which was not conceivable in PBT, may better depict the continually growing construct of language usage and hence boost task authenticity. In this regard, test designers should investigate new multimodal and interactive activities, their roles in various platforms, and their differential influences on the performance of test takers with various characteristics and for various test goals.

In particular, two items including feeling comfortable reading articles on a computer, and forgetting the time while working on a computer significantly and positively correlated with participants' CBT performance. This finding is supported by Chan et al. (2018) who also found that participants who had frequent access to computers and enjoyed working on computers tended to perform better than those who did not. One possible explanation for the finding of this study is that students who can concentrate on analyzing and extracting specific information from the text are more likely to achieve better results in the CBT reading comprehension test.

Test takers' attitudes towards CBT (RQ3)

The attitude survey revealed that most test takers favored CBT, which was consistent with the results reported in previous studies by Higgins et al. (2005) and Khoshsima et al. (2017). Responses for question 3 in the interview showed that some group 1 test takers changed their mode preference after the test from negative to positive. Given that the present study found no relationship between computer familiarity and test takers' scores on CBT of TEM 4, this finding might suggest that the participants' initial negative attitude towards CBT was due to their lack of familiarity with CBT rather than a low level of computer familiarity.

Additionally, this study also found that different anxiety levels might cause some differences in performance between the two modes. Studies showed that students with a lower comfort level with computers may experience greater test anxiety, resulting in lower scores of CBT (Hartono, 2019). Some other studies also showed that computer anxiety is not reliant on the degree of computer experience (e.g., Rabiu et al., 2020; Zheng & Cheng, 2018). As Douglas (2013, p. 2) stated, "we must define the language construct to include appropriate technology in light of the target situation and test purpose." In the case of a comfortable environment, test takers may develop a relatively low level of test anxiety and high efficiency in performing the tasks and therefore may be able to achieve the best test performances.

Conclusions

In response to the increasing adoption of computer technology in China, this study investigated the comparability of CBT and PBT in TEM 4 and two learner factors (computer familiarity and mode preference) which might have affected the test performance. MANOVA analysis revealed no significant difference in test scores between the modes, and the computer familiarity survey showed that participants were generally familiar with using a computer. Subsequently, Pearson correlation analysis revealed that factors including the comfort level of reading articles on a computer (item 5) and forgetting the time when using a computer (item 9) were positively correlated with CBT performance. The attitude survey demonstrated that half of the participants held a positive attitude towards CBT. Thematic analysis of the interview data showed that students preferring CBT claimed that they had an innovative test experience and a comfortable test environment, which have increased their efficiency. As for those not preferring CBT, they stated that the mode increased their test anxiety and they were distracted by the noisy test environment, and they also found it difficult to analyze reading passages on a computer.

Appendix A

Computer familiarity questionnaire

Put a (✓) under the number that describes your perception.

Number	Questions	5-Always 4-Very often 3-Often 2-Once a while 1-Never
1	How often do you use a computer?	5 4 3 2 1
2	How often do you use a computer for...? a) Entertainment (e.g., games, online communications, painting)? b) Studying (e.g., reading, text analysis, word processing)?	5 4 3 2 1 5 4 3 2 1
		5-Very comfortable 4-Quite comfortable 3-Comfortable 2-Quite uncomfortable 1-Very uncomfortable
3	How comfortable are you using a computer in general?	5 4 3 2 1
4	How comfortable are you using a computer to read an article?	5 4 3 2 1
5	How do you feel about using the keyboard (typing)?	5 4 3 2 1
	To what extent do you agree with the following statements?	5-Strongly agree 4-Mostly agree 3-Neutral 2-Mostly disagree 1-Strongly disagree
6	I think taking exams on the computer is an inevitable trend.	5 4 3 2 1
7	I am confident that my ability to use a computer is enough for me to take the computer-based language assessments.	5 4 3 2 1
8	I think taking exams on a computer is interesting.	5 4 3 2 1
9	When I am working with the computer, I always forget about the time.	5 4 3 2 1
		5-Excellent 4-Good 3-Fair 2-Poor 1-Very poor
10	If you compare yourself with other students, how would you rate your familiarity with using a computer?	5 4 3 2 1

Appendix B

Attitude survey

Section A

Put a (✓) under the number that describes your perception.

Number	Questions	Strongly agree-5	Agree-4	Neutral-3	Disagree-2	Strongly disagree-1
A	I found the reading passages easy to navigate through in CBT of TEM 4.					
B	In CBT of TEM 4, I can edit my essay more easily than with pencil and paper.					
C	I find it easy to listen and type at the same time in the listening section of CBT of TEM 4.					
D	I was not distracted by the typing sound made by other test takers.					

Appendix B *(Continued)*

Number	Questions	Strongly agree-5	Agree-4	Neutral-3	Disagree-2	Strongly disagree-1
E	Taking CBT did not increase my test anxiety.					
F	In comparison with PBT, I feel more relaxed while taking CBT.					

Section B

G. What is your overall attitude towards the computer-based version of TEM 4? ( )

a. Positive

b. Negative

c. Neutral

Appendix C

Interview questions

1. What was your perception of computer-based tests before taking CBT of TEM 4?
2. What was your perception of computer-based tests after taking CBT of TEM 4?
3. Has your attitude towards CBT of TEM 4 changed? Please explain why your attitude has changed. You are expected to include the features of CBT of TEM 4 you like and dislike.

Appendix D

Rubric

Band score (points)	Description
2	Ideas are not relevant to the theme. Cohesion is poor. Language errors are very frequent and serious.
6	Ideas are slightly relevant to the theme. Cohesion is poor. Language errors are frequent and serious.
10	Ideas are relevant to the theme. Cohesion is a little poor. Language errors are serious.
14	Ideas are clear and relevant to the theme. The text is cohesive with small language errors.
18	Ideas are closely relevant to the them. The text is very cohesive. language errors are minor.

Appendix E

Table 8 Descriptive statistics by sections on CBT and PBT (NCBT = 46, NPBT = 46)

Section	Full mark	Group	*M*	*SD*	Kurtosis	Skewness
Listening	30	CBT	22.7	3.46	-0.66	-0.89
		PBT	22.8	3.77	-0.51	-0.64
Reading	15	CBT	11.3	2.36	-0.32	-1.22
		PBT	10.9	2.26	-0.58	-1.09
Writing Final	20	CBT	13.2	3.05	0.39	-1.60
		PBT	13.1	3.01	0.61	-1.49
Total Score	65	CBT	47.2	8.12	-0.45	-0.91
		PBT	46.9	8.45	-0.31	-0.84

Table 9 Levene's test of equality of error variances

		Levene statistic	df1	df2	Sig.
Listening 30	Based on mean	0.129	1	90	0.720
	Based on median	0.033	1	90	0.857
	Based on median and with adjusted df	0.033	1	82.454	0.857
	Based on trimmed mean	0.127	1	90	0.723
Reading 15	Based on mean	0.009	1	90	0.923
	Based on median	0.000	1	90	1.000
	Based on median and with adjusted df	0.000	1	87.534	1.000
	Based on trimmed mean	0.014	1	90	0.906
Writing final 20	Based on mean	0.128	1	90	0.722
	Based on median	0.107	1	90	0.745
	Based on median and with adjusted df	0.107	1	89.653	0.745
	Based on trimmed mean	0.133	1	90	0.716

Design: Intercept + Category

Table 10 Multivariate tests

Effect		Value	*F*	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	0.983	1716.891[a]	3.000	88.000	0.000
	Wilks' lambda	0.017	1716.891[a]	3.000	88.000	0.000
	Hotelling's trace	58.530	1716.891[a]	3.000	88.000	0.000
	Roy's largest root	58.530	1716.891[a]	3.000	88.000	0.000
Category	Pillai's trace	0.022	.657[a]	3.000	88.000	0.581
	Wilks' lambda	0.978	.657[a]	3.000	88.000	0.581
	Hotelling's trace	0.022	.657[a]	3.000	88.000	0.581
	Roy's largest root	0.022	.657[a]	3.000	88.000	0.581

Design: Intercept + Category

[a]Exact statistic

Abbreviations

CBT: Computer-based testing; CET: College English Test; EFL: English as a Foreign Language; GRE: Graduate Record Examinations; TAE: Test of Academic English; TEM: Test for English Majors; TEM 4: Test for English Majors-Band 4; NCEE: National College Entrance Examination; MANOVA: Multivariate analysis of variance; MOE: Ministry of Education; PBT: Paper-based testing; PET: Preliminary English Test; SES: Socioeconomic status

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40468-021-00147-0.

Additional file 1.

Acknowledgements

The authors would like to thank all participant learners and instructors who participated in this study.

Authors' contributions

WY contributed to designing the study, acquiring and analyzing data, and writing the manuscript. NI mentored WY through the whole research process and assisted WY to develop the research topic, questions, methodology, and milestones required for successful completion. The author(s) read and approved the final manuscript.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declaration

Competing interests

The authors declare that they have no competing interests.

Received: 3 August 2021 Accepted: 4 November 2021
Published online: 25 November 2021

References

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

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

Bai, Y. (2020). The relationship of test takers' learning motivation, attitudes towards the actual test use and test performance of the College English Test in China. *Language Testing in Asia*, *10*(1). https://doi.org/10.1186/s40468-020-00108-z.

Banerjee, J. (2003). The TOEFL CBT (Computer-based test). *Language Testing*, *20*(1), 111–123. https://doi.org/10.1191/0265532203lt246xx.

Battle, J., & Lewis, M. (2002). The increasing significance of class: The relative effects of race and socioeconomic status on academic achievement. *Journal of Poverty*, *6*(2), 21–35. https://doi.org/10.1300/j134v06n02_02.

Berninger, V. W., Abbott, R. D., Abbott, S. P., Graham, S., & Richards, T. (2002). Writing and reading. *Journal of Learning Disabilities*, *35*(1), 39–56. https://doi.org/10.1177/002221940203500104.

Boo, J., & Vispoel, W. (2012). Computer versus paper-and-pencil assessment of educational development: A comparison of psychometric features and examinee preferences. *Psychological Reports*, *111*(2), 443–460. https://doi.org/10.2466/10.03.11.pr0.111.5.443-460.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa.

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. (2001). *Computer applications in second language acquisition: foundations for teaching, testing and research*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524681.

Chapelle, C. A. (2007). Technology and second language acquisition. *Annual Review of Applied Linguistics.*, *27*, 98–114. https://doi.org/10.1017/S0267190508070050.

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

Coniam, D. (2006). Evaluating computer-based and paper-based versions of an English-language listening test. *ReCALL*, *18*(2), 193–211. https://doi.org/10.1017/s0958344006000425.

Douglas, D. (2013). Technology and language testing. In C. A. Chapelle (Ed.), *The Encyclopedia of applied linguistics* Wiley-Blackwell.

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), 128. https://doi.org/10.3991/ijet.v14i07.10175.

Escudier, M. P., Newton, T. J., Cox, M. J., Reynolds, P. A., & Odell, E. W. (2011). University students' attainment and perceptions of computer delivered assessment; a comparison between computer-based and traditional tests in a "high-stakes" examination. *Journal of Computer Assisted Learning*, *27*(5), 440–447. https://doi.org/10.1111/j.1365-2729.2011.00409.x.

Fan, J., Frost, K., & Liu, B. (2020). Teachers' involvement in high-stakes language assessment reforms: The case of Test for English Majors (TEM) in China. *Studies in Educational Evaluation*, *66*, 100898. https://doi.org/10.1016/j.stueduc.2020.100898.

Florit, E., Roch, M., & Levorato, M. C. (2013). The relation between listening comprehension of text and sentences in preschoolers: Specific or mediated by lower and higher level components? *Applied Psycholinguistics*, *34*(2), 395–415. https://doi.org/10.1017/S0142716411000749.

Flowers, C., Kim, D. H., 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.

Fulcher, G. (1999). Computerizing an English language placement test. *ELT Journal*, *53*(4), 289–299. https://doi.org/10.1093/elt/53.4.289.

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

Hartono, D. A. (2019). Investigating the relationship between test-taking anxiety and test-takers' performance on the IELTS test. *Script Journal: Journal of Linguistic and English Teaching*, *4*(1), 1. https://doi.org/10.24903/sj.v4i1.282.

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), 1-36.

Hosseini, M. (2017). Replacing paper-based testing with an alternative for the assessment of Iranian undergraduate students: Administration mode effect on testing performance. *International Journal of Language and Linguistics*, *5*(3), 78. https://doi.org/10.11648/j.ijll.20170503.13.

Hosseini, M., Abidin, M. J. Z., Kamarzarrin, H., & Khaledian, M. (2013). The investigation of difference between PPT and CBT results of EFL learners in Iran: Computer familiarity and test performance in CBT. *International Letters of Social and Humanistic Sciences*, *11*, 66–75. https://doi.org/10.18052/www.scipress.com/ilshs.11.66.

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.

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

Jiao, H., & Wang, S. (2010). A multifaceted approach to investigating the equivalence of computer-based and paper-and-pencil assessments: An example of reading diagnostics. *International Journal of Learning Technology*, *5*(3), 264. https://doi.org/10.1504/ijlt.2010.037307.

Jin, Y. (2010). The place of language testing and assessment in the professional preparation of foreign language teachers in China. *Language Testing*, *27*(4), 555–584. https://doi.org/10.1177/0265532209351431.

Jin, Y., & Fan, J. (2011). Test for English Majors (TEM) in China. *Language Testing*, *28*(4), 589–596. https://doi.org/10.1177/0265532211414852.

Jin, Y., & Yan, M. (2017). Computer literacy and the construct validity of a high-stakes computer-based writing assessment. *Language Assessment Quarterly*, *14*(2), 101–119. https://doi.org/10.1080/15434303.2016.1261293.

Juel, C., Griffith, P. L., & Gough, P. B. (1986). Acquisition of literacy: A longitudinal study of children in first and second grade. *Journal of Educational Psychology*, *78*(4), 243–255. https://doi.org/10.1037/0022-0663.78.4.243.

Khoshsima, H., Hosseini, M., & Toroujeni, S. M. H. (2017). Cross-Mode comparability of computer-based testing (CBT) versus paper-pencil based testing (PPT): An investigation of testing administration mode among iranian intermediate EFL learners. *English Language Teaching*, *10*(2), 23. https://doi.org/10.5539/elt.v10n2p23.

Khoshsima, H., Toroujeni, S. M. H., 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 http://www.tewtjournal.org.

Kim, H. R., Bowles, M., Yan, X., & Chung, S. J. (2018). Examining the comparability between paper- and computer-based versions of an integrated writing placement test. *Assessing Writing*, *36*, 49–62. https://doi.org/10.1016/j.asw.2018.03.006.

Kim, Y.-S., Al Otaiba, S., Puranik, C., Folsom, J. S., Greulich, L., & Wagner, R. K. (2011). Componential skills of beginning writing: An exploratory study. *Learning and Individual Differences*, *21*(5), 517–525. https://doi.org/10.1016/j.lindif.2011.06.004.

Kim, Y.-S., & Phillips, B. (2014). Cognitive correlates of listening comprehension. *Reading Research Quarterly*, *49*(3), 269–281. https://doi.org/10.1002/rrq.74.

Kim, Y.-S. G. (2016). Direct and mediated effects of language and cognitive skills on comprehension or oral narrative texts (listening comprehension) for children. *Journal of Experimental Child Psychology*, *141*, 101–120. https://doi.org/10.1016/j.jecp.2015.08.003.

Kolen, M. J. (1999). Threats to score comparability with applications to performance assessments and computerized adaptive tests. *Educational Assessment*, *6*(2), 73–96. https://doi.org/10.1207/s15326977ea0602_01.

Kunnan, A. J. (2015). *Language testing and assessment*. London, England: Routledge.

Kunnan, A. J., & Carr, N. (2017). A comparability study between the General English Proficiency Test- Advanced and the Internet-Based Test of English as a Foreign Language. *Language Testing in Asia*, *7*(1). https://doi.org/10.1186/s40468-017-0048-x.

Li, X. (2019). Language Testing in China: Past and Future. *English Language Teaching*, *12*(12), 67. https://doi.org/10.5539/elt.v12n12p67.

Mohammadi, M., & Barzgaran, M. (2012). Comparability of computer-based and paper-based versions of writing section of PET in Iranian EFL context. *Journal of Foreign Language Teaching and Translation Studies*, *1*(2), 1–20.

Odo, D. M. (2012). Computer familiarity and test performance on a computer-based cloze ESL reading assessment. *Teaching English with Technology*, *12*(3), 18–35.

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.

Patel, A. B., & Laher, S. (2011). The influence of mode of test administration on test performance. *Journal of Psychology in Africa*, *21*(1), 139–144. https://doi.org/10.1080/14330237.2011.10820440.

Prisacari, A. A., & Danielson, J. (2017). Computer-based versus paper-based testing: Investigating testing mode with cognitive load and scratch paper use. *Computers in Human Behavior*, *77*, 1–10. https://doi.org/10.1016/j.chb.2017.07.044.

Qian, D. D., & Cumming, A. (2017). Researching English language assessment in China: Focusing on high-stakes testing. *Language Assessment Quarterly*, *14*(2), 97–100. https://doi.org/10.1080/15434303.2017.1295969.

Rabiu, N., Kehinde, A., Amuda, H. O., & Kadiri, K. K. (2020). University of Ilorin undergraduate students' perceptions of the usefulness and challenges regarding computer-based testing. *Mousaion: South African Journal of Information Studies*, *37*(4). https://doi.org/10.25159/2663-659x/7305.

Randall, J., Sireci, S., Li, X., & Kaira, L. (2012). Evaluating the comparability of paper- and computer-based science tests across sex and SES subgroups. *Educational Measurement: Issues and Practice*, *31*(4), 2–12. https://doi.org/10.1111/j.1745-3992.2012.00252.x.

Sawaki, Y. (2012). Technology in language testing. In G. Fulcher, & F. Davidson (Eds.), *The Routledge handbook of language testing*, (pp. 426–437). Routledge.

Srivastava, P., & Gray, S. (2012). Computer-based and paper-based reading comprehension in adolescents with typical language development and language-learning disabilities. *Language, Speech, and Hearing Services in Schools*, *43*(4), 424–437. https://doi.org/10.1044/0161-1461(2012/10-0108).

Tajuddin, E. S., & Mohamad, F. S. (2019). Paper versus screen: Impact on reading comprehension and speed. *Indonesian Journal of Education Methods Development*, *3*(2). https://doi.org/10.21070/ijemd.v3i2.20.

Tompkins, V., Guo, Y., & Justice, L. M. (2013). Inference generation, story comprehension, and language in the preschool years. *Reading and Writing: An Interdisciplinary Journal*, *26*(3), 403–429. https://doi.org/10.1007/s11145-012-9374-7.

Uke, W. A. S. (2017). Students' perception towards national examination 2017: Computer-based test or paper-based test. *Mediterranean Journal of Social Sciences*, *8*(4-1), 139–143. https://doi.org/10.2478/mjss-2018-0083.

Walpole, M. (2003). Socioeconomic status and college: How SES affects college experiences and outcomes. *The Review of Higher Education*, *27*(1), 45–73. https://doi.org/10.1353/rhe.2003.0044.

Wang, S., Hong, J., Young, M. J., Brooks, T., & Olson, J. (2007). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments. *Educational and Psychological Measurement*, *68*(1), 5–24. https://doi.org/10.1177/0013164407305592.

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://doi.org/10.1111/j.1745-3984.2001.tb01115.x.

Weir, C. J., O'Sullivan, B., Yan, J., & Bax, S. (2007). Does the computer make a difference? Reaction of participants to a computer-based versus a traditional handwritten form of the IELTS writing component: effects and impact. *IELTS Research Report*, *7*, 1–37 IELTS Australia, Canberra and British Council, London.

Yan, X., & Fan, J. (2020). "Am I qualified to be a language tester?": Understanding the development of language assessment literacy across three stakeholder groups. *Language Testing*, *38*(2), 219–246. https://doi.org/10.1177/0265532220929924.

Yao, D. (2019). A comparative study of test takers' performance on computer-based test and paper-based test across different CEFR levels. *English Language Teaching*, *13*(1), 124. https://doi.org/10.5539/elt.v13n1p124.

Yeom, S., & Jun, H. (2020). Young Korean EFL learners' reading and test-taking strategies in a paper and a computer-based reading comprehension tests. *Language Assessment Quarterly*, 1–18. https://doi.org/10.1080/15434303.2020.1731753.

Yu, G., & Jin, Y. (2014). English language assessment in China: policies, practices and impacts. *Assessment in Education: Principles, Policy & Practice*, *21*(3), 245–250. https://doi.org/10.1080/0969594x.2014.937936.

Yu, G., & Zhang, J. (2017). Computer-based English language testing in China: Present and future. *Language Assessment Quarterly*, *14*(2), 177–188. https://doi.org/10.1080/15434303.2017.1303704.

Zhang, Q. (2021). Impacts of World Englishes on local standardized language proficiency testing in the Expanding Circle. *English Today*, 1–17. https://doi.org/10.1017/s0266078421000158.

Zheng, Y., & Cheng, L. (2018). How does anxiety influence language performance? From the perspectives of foreign language classroom anxiety and cognitive test anxiety. *Language Testing in Asia*, *8*(1). https://doi.org/10.1186/s40468-018-0065-4.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.