Computerized testing in reading comprehension skill: investigating score interchangeability, item review, age and gender stereotypes, ICT literacy and computer attitudes

Seyyed Morteza Hashemi Toroujeni¹

Received: 4 December 2020 / Accepted: 21 April 2021 / Published online: 3 August 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

Score interchangeability of Computerized Fixed-Length Linear Testing (henceforth CFLT) and Paper-and-Pencil-Based Testing (henceforth PPBT) has become a controversial issue over the last decade when technology has meaningfully restructured methods of the educational assessment. Given this controversy, various testing guidelines published on computerized testing may be used to investigate the interchangeability of CFLT and PPBT mean scores to corroborate if test takers' testing performance is influenced by the effects of testing administration mode; specifically, if validity and reliability of two versions of the same test are affected. This research was conducted to probe not only score interchangeability across testing modes but also to explore the role of age and gender stereotypes, item review, ICT literacy and attitudes towards computer use as moderator variables in test takers' reading achievement in CFLT. Fifty-eight EFL learners homogeneous in both general English and reading skills assigned into one testing group participated in this study. Three different versions of TOEFL reading comprehension test, Computer Attitude Scale (CAS), and ICT literacy Scale of TOEFL Examinees were used in this crossover quasi-controlled empirical study with a common-person and pretest–posttest design to collect data. The findings demonstrated that although the reading scores of test takers were interchangeable in both CFLT and PPBT versions regarding testing administration modes, they were different regarding item review. Furthermore, no significant interaction was found between age, gender, and ICT literacy and CFLT performance. However, attitudes towards the use of computer led to a significant change in testing achievement on CFLT.

✉ Seyyed Morteza Hashemi Toroujeni
hashemi.seyyedmorteza@gmail.com

Extended author information available on the last page of the article

Keywords Computer-based testing · Score interchangeability · Testing administration mode effect · Item review · ICT literacy · Computer attitudes

1 Introduction

Use of computer as an integral (Stole et al., 2020) and ubiquitous part of every society (Mullis, Martin, Foy, and Hooper, 2017) raised testing productivity by introducing efficacious tools that have some advantages (Shraim, 2019) such as evaluating responses and giving proper prompt feedback in an intelligent environment (Kilickaya, 2013). Popularity of using computer to improve testing approaches and solve large-scale assessment problems that is increasingly growing in recent years (Sawada, 2019) caused entirely practical reconstructions in assessment procedures (Khoshshima and Hashemi Toroujeni, 2017h). In particular, Covid-19 identification late 2019 and health crisis caused by Coronavirus outbreak (WHO, 2020) all over the world (Karim & Hasan, 2020) led education systems of especially Asian developing countries to face their greatest challenge because their technology infrastructure was not sufficient (Retnawati, 2015) for digitalized education, and investments on technology (ADB, 2018; Sawada, 2019) is still developing in these countries for a widespread use of computer-based teaching, learning and testing as the new approaches (Karim & Hasan, 2020). Due to COVID-19 pandemic, face to face learning has been discontinued and replaced by different ways of learning straight through technology (Alghammas, 2020) in most parts of the world especially in developing Asian countries such as Iran. Consequently, shutting universities, schools, and institutes, and returning to educating by shifting from conventional PPBT to computerized version of testing to evaluate learners has happened again as a prevailing approach for assessment (Ary et al., 2018) into the current digitalization of assessment era (Ben-Yehudah & Eshet-Alkalai, 2020) in Iran in April 2021 after reopening in September 2020.

Then, transition from paper-and-pencil-based conventional testing (PPBT) to Computerized Fixed-Length Linear Testing (CFLT), which has been usually encouraged as an innovative approach and alternative assessment (Shute & Rahimi, 2017; Thurlow et al., 2010) due to its financial advantages, cost-effective administration (Khoshshima, Hahemi Toroujeni, Thompson and Ebrahimi, 2019), and its capability to measure and promote cognitive functions of brain (Galindo-Aldana, Meza-Kubo, Ledesma-Amaya, Galarza-Del-Angel, Padilla-Lopez and Moran, 2018) since before coronavirus outbreak, is currently a necessity in education. For example, in spite of the low correlation of conventional academic section of International English Language Testing System (IELTS) exam with its converted CFLT counterpart discovered in 2005, test scores obtained from two versions have been used interchangeably (Blackhurst, 2005) to make testing administration more economic (Schroeders & Wilhelm, 2011), and to enjoy its administrative advantages (Yu & Zhang, 2017) such as huge long-term decrease in testing cycle costs from item generation, test administration, scoring process and reporting results immediately after the termination of CFLT. Although interchangeability of scores from two CFLT and PPBT modes has been investigated in testing domain and general English in several studies, rising prevalence of technology in different domains of education such as assessment and much more familiarity of teachers

and learners (Katz & Elliot, 2016) with technology demand continuing this field of research to realize how shifting from PPBT to CFLT may have impacts on achievement of test takers in reading comprehension.

In Iran, while the importance of computer is recognized in developing different models of testing due to the advantages such as direct scoring algorithm, productive and well-organized administration of tests (Khoshshima et al., 2019), more efficient and manageable scheduling (Kumar, 2013; Shraim, 2019), fast scoring and result reporting (Oz & Ozturan, 2018), immediate automaticity of feedback (Stole et al., 2020), and greater accuracy, enjoyment (Boeve et al., 2015) and security (Burr, Chatterjee, Gibson, Coombes and Wilkinson, 2016), there are still concerns of validity or reliability challenges that CFLT may cause for assessment efficiency. Recognition of the importance, enthusiastic reception and preference for CFLT do not guarantee validity of CFLT. For example, just one test taker out of the total number of 319,000 test takers who took a wide variety of CFLT version during seven years preferred PPBT version over CFLT, and the other test takers endorsed CFLT (Bugbee, 1996). A CFLT version that is highly endorsed by test takers may not be valid, reliable and equivalent to its PPBT counterpart. PPBT and CFLT equivalency means that validity (the degree to which a specific test measures exactly what it claims to measure) (Stobart, 2012) and reliability (the degree to which a test measures consistently and stably what it claims to measure) (Scheerens, Glas, and Thomas, 2005, p. 93) of a test are not violated as a result of transition (Khoshshima and Hashemi Toroujeni, 2017h). Dogan et al., (2020) believe that because transition of PPBT to CFLT is growing due to the universal increasing use of personal computers, there is no controversy against two versions equivalency in most of assessments. On the other hand, they state that validity and reliability issues regarding measurements are important (Dogan et al., 2020). It should be noted that comparability studies are done to explore whether results obtained from various testing items, materials and procedures, or different administration modes can be used interchangeably. If a test is implemented twice in different times or parallel versions and the same results are obtained, the test, two versions, and the results are considered reliable, equivalent and interchangeable, respectively (Brown & Abeywickrama, 2010; Oz & Ozturan, 2018). According to Hughes (2003), if reliability of a test is satisfied, validity as the most significant criteria for assessment (AERA, APA and NCME 2014) is then assured. Assuring reliability and validity of a test as psychometric values increases transparency and accuracy of evaluation (Tavakol & Dennick, 2011). As reported by Carpenter and Alloway (2018), CFLT and PPBT are not equivalently reliable unless they yield equivalent scores. If a test is reliable, test-takers get almost similar scores regardless of when the test is implemented, and what version of the test is administered.

In response to increasing Covid-19 pandemic and shutting down schools (Hashemi Toroujeni et al., *In Press*), Iran Ministry of Education that has faced its greatest learning challenge during recent decades launched its educational network of student with releasing communication and educational software known as SHAD for more than 14 million students since late 2020 due to the vital role of ICT in education (Leino, 2014) so that students could continue education during taking staying home strategy (Sintema, 2020) through social distancing and homeschooling (Pokhrel & Chhetri, 2021). Iran Ministry of Education announced that students had to follow learning via SHAD, and take their exams through computer or mobile mode until the end of the current educational

calendar (August 2021). Furthermore, in universities, remote learning and assessment through free software web conferencing systems such as BigBlueButton is being used, and the exams are usually conducted through computer, because some features of software may not be shown in mobile. Score interchangeability between CFLT and PPBT as the controversial issue in the last decade (Sangmeister, 2017) in Computer-Based Testing field should be investigated empirically through studying psychometric properties (Burr et al., 2016) as the prerequisite of replacing CFLT with PPBT (Khoshsima et al., 2019; Rausch, Seifried, Wuttke, Kogler and Brandt, 2016). Therefore, since authenticating and sustaining validity and reliability of measurement is essential for replacing CFLT with PPBT, the current study concentrated on whether validity and reliability measurement would be violated by changing testing administration mode and whether scores received from PPBT and CFLT would be interchangeable or equivalent (TEA, 2008).

Some years ago, although it was believed that CFLT would completely replace PPBT (Garcia Laborda, Magal Royo, and Enriquez Carrasco, 2010; Tahmasebi & Rahimi, 2013) due to the huge popularity of CFLT in education (Hardcastle, Hermann-Abell, and DeBoer, 2017), both modes of testing administration still co-exist and are delivered together by many institutes and educational organizations such as ETS for TOEFL to assess progress in educational attainments. However, unanswered questions still remain on whether received scores from CFLT are comparable to the scores generated in PPBT and whether two sets of scores are equivalent measures of test takers' performance (Hardcastle et al., 2017).

When the same or similar test is implemented in its alternative mode, and received scores demonstrate that test takers show the same level of proficiency, then scores are considered reliable. The alternative versions of tests should produce sustainable valid and reliable measures of intended proficiency (Newhouse & Cooper, 2013). According to the guidelines published by American Educational Research Association (AERA), if more than one way of different ways of implementing a test is used, scores received from the ways should be interchangeable (AERA, 2014). Then, equivalency across CFLT and PPBT delivery modes in education is of great importance because assessment of academic progress is usually done through paper and computer (Blazer, 2010) across different times (Csapo, Ainley, Bennett, Latour, and Law, 2012) especially during the Covid-19 pandemic since late 2019. Furthermore, in the age of technologizing assessment (Ary et al., 2018), teachers are capable and intelligent enough to create their own CFLT versions to assess their students' attainments, and consequently to make instructional decisions (Hensley, 2015). This is the main reason leads some researchers in middle-eastern countries such as Iran, Japan, Hong-Kong, China, Thailand, Turkey, Saudi Arabia, Malaysia, and Jordan (Hashemi Toroujeni, Thompson, and Faghihi, *In Press*) to investigate whether test-takers' scores are equivalent across two test versions (Alakyleh, 2018). Therefore, scores across two delivery modes or across different times need to be interchangeable or equivalent. CFLT and PPBT versions of a test are called equivalent, valid and reliable if the same content covering the same skills generate similar scores.

Some studies report score interchangeability and no statistically significant difference between paper-based and computerized tests (Hashemi Toroujeni et al., *in press*;

Khoshshima and Hashemi Toroujeni, 2017h; Prisacari & Danielson, 2017; Register-Mihalik et al., 2012). Although Ebrahimi, Hashemi Toroujeni and Shahbazi (2019), Hermena et al. (2017), Khoshshima et al. (2019), Khoshshima and Hashemi Toroujeni (2017h), Porion et al. (2016) indicate that two identical computer-based and paper-based tests may result in the same scores; some others reveal different test results (Emerson & MacKay, 2011; Galindo-Aldana et al., 2018; Jerrim, 2016; Jerrim et al., 2018; Kim & Kim, 2013; Washburn, Herman, Stewart, 2017) especially in reading comprehension skill (Clinton, 2019; Delgado et al., 2018; Stole et al., 2020) due to the “Testing Mode Effect.” Such empirical findings help testing practitioners decide whether to replace computer-based testing with its identical paper-based test. However, researchers have not yet reached an agreement on a comprehensive theoretical explanation for testing mode effect. Given these conflicting findings, the researchers consider that the issue of testing mode effects on the equivalency of data attained from two CFLT and PPBT presentation modes needs attention and prompt investigation.

Converting conventional PPBT version of a test into its computerized counterpart might become problematic when considering reliability and validity. Constructing reliable and valid tests are the main concerns in utilizing CFLT. Then, a CFLT whose psychometric properties (Burr, et al., 2016) and validity and reliability (Johnson & Green, 2006) are matched with its conventional counterpart can assist test takers to attain their accurate achievement. Evaluation of validity and reliability is, therefore, the reason for doing many of comparability studies between CFLT and PPBT (Al-Amri, 2007; Hashemi Toroujeni, 2016). A test is reliable when it regularly measures what it is expected to measure by producing stable and constant scores on two testing occasions. In other words, a test can be considered reliable when constant similar results or scores are repeated under the same conditions (Vansickle, 2015). Therefore, it is important to examine reliability and validity of a computerized test by conducting a comparability study, particularly, in a local context, to establish any testing mode effects that result from converting a conventional test into its computerized counterpart. One of the major goals pursued in comparability studies is to examine interchangeability of test scores across different modes of administration. To achieve this goal, test items should be presented uniformly across two modes.

However, we can expect the same or evenly matched scores in both modes of administration when we administer two identical tests covering similar materials; the more identical and interchangeable the scores of two modes, the more reliable and equivalent the test is in a consistent manner (Smolinsky et al., 2020). When tasks are moved from pen and paper to computer, equivalence is often assumed, but this is not necessarily the case. For example, even if paper version has been shown to be valid and reliable, computer version may not exhibit similar characteristics. If equivalence is required, then it needs to be established (Noyes & Garland, 2008).

Since test takers’ achievement on CFLT depends on both their proficiency in testing materials, testing skills, their computer skills (Zhu & Aryadoust, 2020) and other commonly seen characteristics such as ICT literacy, attitudes, the researchers of the current study made a decision to investigate aforementioned characteristics’ roles in test takers’ language proficiency in CAT version. Some comparability studies investigating two testing administration modes have been done across some characteristics such as racial-ethnic groups, age, gender, and types of items (Carpenter & Alloway,

2018; Horne, 2007; Piaw, 2012). Then, in addition to the technical concepts of scores interchangeability, ICT literacy and computer attitude, item review, age, and gender stereotypes were investigated in the current study as these are major highly influencing characteristics in respect of a test taker's performance. Therefore, to study testing administration mode variable, the research data gathered in a pre and post-test design were analyzed to discover what variables may be considered as moderators regarding testing mode effect.

2 Literature review

Covid-19 identification late 2019 and health crisis caused by Coronavirus outbreak (WHO, 2020) all over the world (Karim & Hasan, 2020) led especially education systems of Asian developing countries to face their greatest challenge because their technology infrastructure was not sufficient (Retnawati, 2015) for digitalized education. Investment on technology (ADB, 2017; ADB, 2018; Sawada, 2019) is still developing in these countries for a widespread use of computer-based teaching, learning and testing as the new approaches (Karim & Hasan, 2020) in education. In some contexts, since computerized versions of tests are available, users have a choice between taking the test in either mode. Converting paper and pencil assessment into computerized version often requires that the computerized version be comparable and equivalent to the conventional paper and pencil one and the scores obtained from two identical tests approximate to each other. To consider a test reliable and valid, score interchangeability is required for test takers who are administered two identical tests in either mode (Bartram & Hambleton, 2016). Although some studies demonstrated similar results asserting that substantial testing mode effect was seen in speeded tests (Pomplum, Frey and Becker, 2002), some others found that mode effects were not observed on non-speeded tests with a short answer format (Wang et al. 2007). Some researchers achieved lower scores on CFLTs (Chen et al. 2011) while others received higher scores on CFLTs (Clariana & Wallace, 2002; Pomplum et al., 2002). In some studies, test takers outperformed on PPBT rather than CFLT (Carpenter & Alloway, 2018; Hosseini et al., 2014), or no testing administration mode effect was found (Jeong, 2012; Karay et al., 2015; Meyer et al., 2016; Prisacari & Danielson, 2017). Although these results cannot be described as decisive, there is a growing tendency to suggest that two CFLT and PPBT versions are expected to be equivalent across two presentation modes (Alakyleh, 2018; Ebrahimi et al, 2019; Khoshsima et al, 2019; Wang & Shin, 2010). Converting PPBT into CFLT and studying mode effect on testing performance should be done through carefully well-organized empirical investigations. Conducting these kinds of comparability investigations help test developers to find out if the scores obtained from computerized tests remain valid and that students are not disadvantaged by taking CFLT.

During the global COVID-19 enforced lockdowns and homeschooling (Pokhrel & Chhetri, 2021) when about half of the world's population (Sandford, 2020) and more than 98% of learners (United Nations, 2020) were affected by the coronavirus outbreak, in-person learning was shifted to remote education (Pokhrel & Chhetri, 2021) and digital learning (Dhawan, 2020) through computer and mobile modes of presentation (Hashemi Toroujeni, et al., In Press). Consequently, due to the synchronizing

remarkably arising prevalence and availability of ICT (Gnambs, 2021) and technological advancements (Siddiq & Scherer, 2019) with the ubiquity of computers and smartphones use (Mullis et al., 2017) in daily lives of learners (Daghan, 2017) in last years (Ebrahimi et al., 2019; Khoshshima et al., 2019; Garcia-Laborda and Alcalde-Penalver, 2018) and in the current homeschooling days at increasing spread of COVID-19 (WHO, 2020; Doyle, 2020), many learners have to swapped reading textbooks on screen and from digital resources (Barzillai & Thomson, 2018; Halamish & Elbaz, 2019).

Since several types of new text forms such as e-books (Bando, Gallego, Gertler and Romero, 2016) are being delivered through digital modes, reading texts on-screen seems inevitable (Hancock, SchmidtDaly, Fanfarelli, Wolfe and Szalma, 2016; Purcell et al., 2013) in educational lives of EFL learners. Furthermore, many assessments (Singer & Alexander, 2017a, 2017b) and reading (Golan et al., 2018) are being done digitally. Therefore, the effect of onscreen mode on the comprehension and achievement of learners necessitates a systematic investigation of differences that might be created in reading comprehension when learners have access to a variety of texts in different paper and onscreen modes. EFL learners are assumed to use different strategies when reading a text on paper or onscreen. For example, they may make a connection between their prior knowledge and the knowledge illustrated in the available text (Singer & Alexander, 2017a, 2017b). Any change in the delivery mode of the text may influence their trying to make this connection and fail to achieve the same performance in two different administration modes.

Replacing reading on screen with paper-based reading raised the concerns of affecting cognitive learning outcomes (Stole et al., 2020) and impairing reading comprehension (Halamish & Elbaz, 2019; Margolin et al. 2013) due to the effect of transitioning presentation mode (Chen et al., 2014; Halamish & Elbaz, 2019). Then, it seems crucial to find out whether learners' reading comprehension achievement is different when they read text in two on-screen and paper modes.

The major motivation behind conducting the current research is the increasingly converting international PPBT assessments into CFLT such as PISA (Backes & Cowan, 2018), TOEFL and IELTS that indicates the importance of CFLT and adds more value to this topic. On-screen versus paper-based reading comprehension and reading across presentation modes have been the topic of some empirical investigations led to inconsistent findings (Porion et al., 2016; Stole et al., 2020) during last years. Although the effect of some text characteristics such as font size and type were investigated (French et al., 2013; Pieger et al., 2016), the controversial issue of reading text from screen (Clinton, 2019) involves a highly related unanswered question about the effect of medium (Halamish & Elbaz, 2019; Singer-Trakhman, Alexander, and Berkowitz, 2019) and mode of text presentation in reading comprehension skill. Then, this study aims at comparing reading achievements of Iranian intermediate EFL learners from computer (CFLT) and paper (PPBT), as well as item review, age and gender differences, ICT literacy and attitudes towards use of computer while reading from screen is reported less charming and delightful (Mangen & Kuiken, 2014) compared to reading from paper.

Existing literature reported conflicting findings on the effect of medium transition on reading comprehension achievement or on the benefit of reading from screen or paper. Some found no or comparable effect (Chen & Catrambone, 2015; Farinosi

et al., 2016; Hermena et al., 2017; Kong et al., 2018; Porion et al., 2016), a few studies found advantages for onscreen (Aydemir et al., 2013; Singer & Alexander, 2017a, 2017b), and some others found advantages for reading from paper (Clinton, 2019; Delgado et al., 2018; Golan et al., 2018; Lenhard, Schroeders, and Lenhard, 2017; Rasmussen, 2015). Heterogeneous findings' collection of studies including hybrid modes of text presentation or testing administration in a variety of contexts and inadequacy of findings in reading comprehension skill in an Asian private EFL context justify this detailed investigation of testing administration mode in reading comprehension and the other moderator factors that might moderate the effect of this administration or presentation mode.

2.1 Critical issues related to the testing administration mode effect

When learners are learning reading skill or taking a reading test through CFLT that is a different experience from conventional learning or testing environment, their learning or testing may be manipulated through text presentation or testing administration mode as the independent variable and some other moderator variables such as item review i.e. the chance that is given to test takers to review their answers, ICT literacy or prior ICT literacy of learners, their negative or positive attitudes towards using digital mode to read the text or take the test, and gender and age differences in reading achievement that may moderate the effect of presentation or administration mode. Item review, i.e., reviewing and changing the answers during a test, is test takers' main concern in the pursuit of improving answers given to a test. Concern of test items modification gains remarkable prominence in multiple-choice tests. The impacts of item review in conventional paper-based tests have been investigated over previous decade (Elliot & Kettler, 2013). Although some findings demonstrated that just a few responses were modified at the item review stage (Revuelta et al. 2003), the importance of this opportunity during a test should not be ignored because it can allow test takers to improve their performance. Item review is an intrinsic feature of paper-based tests, but this option is not usually considered in some models of computerized tests. For example in Computer-Adaptive Tests (CAT) which employ an adaptive form of testing paradigm to tailor each test item to the current abilities of test takers based on their responses, item review can be very problematic since it could have the effect of violating the test validity. Eaves and Smith allowed their paper-based test takers to review the items to modify their responses, but the computerized test takers were not given the opportunity to review items during the test. The findings indicated that this external variable did not affect test performance (Eaves & Smith, 1986). However, in CFLTs that use fixed-length linear algorithm, a significant difference may be possible due to activating item review feature. Since there are only a few studies examining item review in computerized testing, the current researcher imagined it would be important to address this critical issue.

ICT literacy which is usually referred to as digital skill or competence (Ala-Mutka, 2011; Siddiq et al., 2016) plays a vital role in learners' achievement (Pagani et al., 2016) in computerized testing. It was also found that ICT literacy had no

significant effect on test takers' performance and their willingness to take computerized test when two versions of the same test were available (Hashemi Toroujeni et al., In Press). Wallace and Clariana found that learners' ICT literacy was associated with higher post-test performance in computerized testing (in their case, web-based test) (Clariana & Wallace, 2002). Their results showed that learners with lower scores were less familiar with computers. The Florida Department of Education reported that early examinations of the relationship between ICT literacy and test performance showed significant differences (Florida Department of Education, 2006), providing empirical evidence of lower scores of test takers being associated with those who had less experience with computer. However, in some studies, it is stated that there is no relationship between ICT literacy and computerized test performance (Florida Department of Education, 2006). Since some students of the current researcher as the teacher of Iran Ministry of Education declare that their poor ICT literacy and unfamiliarity with computerized mode of testing was the main reason for falling in final exams of schools in last educational year, and their low performance on CFLT is attributed to their poor ICT skills, the researcher came to a decision to consider prior ICT literacy or frequent use of computer as a moderator variable in CFLT in order to demonstrate whether the attained scores are authentic. Investigation on attitudes towards computerized test plays a crucial role in implementing CFLT successfully. Some studies found test takers with positive attitudes towards CFLT (Al-Amri, 2009). Negative or positive attitudes towards use of computer that have a direct relationship with ICT literacy can be influenced by contextual factors such as age, gender, and socioeconomic status. According to Eagly and Shelly, attitude is a positive or negative feeling towards a psychological object (Eagly & Shelly, 1998). In another definition of attitude, Loyd and Gressard identify four components including anxiety, confidence, liking, and usefulness that form attitudes toward computers construct (Loyd & Gressard, 1985). According to their definition, anxiety is a feeling of fear associated with computer use; computer confidence describes the user's ability to use computer or willingness to learn more about it; computer liking refers to the enjoyment associated with working with computers; and computer usefulness is defined as appreciating the efficiency and usefulness of working with computers (Loyd & Gressard, 1985). All of these components form a scale (Computer Attitude Scale (CAS) developed by Loyd and Gressard (Loyd & Gressard, 1985) that examines attitudes towards computers as a whole. In a comparability study conducted by Khoshshima and his colleague, CAS questionnaire was used to evaluate EFL learners' attitudes towards computers (Khoshshima and Hashemi Toroujeni, 2017b) in an academic context. The correlation between attitudes towards computers and results obtained from CFLT and PPBT indicated that test takers performed better on CFLT, even though there was no significant correlation between positive attitudes towards using a computer and testing performance. Al-Amri used some special sections of CAS questionnaire to study learners' attitudes toward computer use. Even though the students showed a high preference for CFLT, his research findings indicated no correlation between learners' attitudes and their performance on CFLT (Al-Amri, 2009). Moreover, the effectiveness difference of testing methods (CFLT vs. PPBT) in terms of age and gender has also been examined in some studies and no statistically significant difference was found (Bennett

et al. 2008) in testing performance, while, in other comparability studies such as (Gallagher, Bridgeman and Cahalan, 2000), a statistically significant difference was found. Terzis and Economides also investigated the relationship between gender and CFLT performance and the trends of female and male test takers towards the features of CFLT (Terzis & Economides, 2011).

The researcher hopes that the current research would add to the good working knowledge of the CFLT version of reading comprehension skill in an EFL context in private education. Therefore, considering both theoretical and pedagogical perspectives to achieve the research objectives mentioned above, the following research questions were addressed:

1. Is there a change in reading comprehension scores for intermediate EFL learners across testing administration modes (PPBT vs. CFLT)? Do (a) gender, (b) age and (c) item review play a role in this change?
2. Do EFL learners' (a) ICT literacy, and (b) attitudes towards the use of computer influence their performance in CFLT version of reading comprehension?

Based on the pedagogical implications of the study, the following null hypotheses are to be tested at the probability level of 0.05.

H01: Administration of CFLT does not affect intermediate EFL learners' reading comprehension. There is no statistically significant correlation between (a) gender, (b) age, and (c) item review and test takers' CFLT scores.

H03: There is no statistically significant correlation of EFL learners' (a) ICT literacy and their (b) attitudes towards use of computer with CFLT performance on reading comprehension

3 Methodologies

3.1 Research design and variables

Quantitative data were gathered from a crossover study in which the six variables were administered sequentially to the same group on three testing occasions (one pre and two post-tests). To answer the research questions and to reject or confirm different sections of the null hypotheses, the first CFLT with an item review option was offered in the second testing session (CFLT1), and the second CFLT was offered with no item review option in the third testing session (CFLT2). Common person design was selected for this research as this enabled the researchers to examine the effect of item review on CFLT performance. In the third testing session (CFLT2), the researchers adopted testing administration mode and item review variables individually. Since the carry-over impact, i.e., the effect of the first treatment or variable on the second one (in our case, the effect of testing administration mode effect on item review variable or vice versa) is considered to be one of the most critical disadvantages in a crossover study, the researchers utilized a third testing session to study the item review impact on test takers' performance separately. As testing administration mode was in the same

format in both post-tests (CFLT1 and CFLT2), the impact of the item review variable on the CFLT testing performance could be measured without the intervention of a third-party treatment or variable. Then, in the second and third testing sessions, the CFLT with item review and CFLT without an item review option were administered to the test takers, respectively. The methodological approach adopted in the current study combined a reading comprehension multiple-choice test (two modes) and two questionnaires to be implemented to a testing group as a critical first step in this comparability study. This experimental design is so powerful in detecting differences especially in smaller samples of test takers to collect and measure the research data before and after applying the treatment(s). Homogenous participants were assigned to a testing group, and the effect of their features such as age, gender, ICT literacy, and computer attitude as well as the testing mode of administration was investigated based on a within-subject group score comparisons (Table 1).

Dependent and independent variables including testing administration mode and item review as well as test takers’ characteristics were critically examined. Dependent variable which was expected to change as the researcher introduced the computerized test and item review option, was the participants’ scores received in a reading comprehension skill exam (administered to the participants in two versions on three testing sessions). It is worth mentioning that the age variable was considered as a dichotomy concept in this research. The participants were divided into two younger (below-30) and older (above-30) classifications; 68.96% (40 out of 58) and 31.04% (18 out of 58) of the participants were categorized as younger and older participants, respectively.

3.2 Participants

The study was conducted at the Adrina Language Academy (ALA) located in a large city in Northern Iran. Although those who attend the EFL classes of Adrina Language Academy take a standardized paper-based placement test to determine the classes and materials appropriate to their English language proficiency level, the researchers preferred to screen the participants and select the most homogeneous ones by administering the TOEFL general proficiency test. The TOEFL test was administered in the autumn of 2019, and 69 students were selected from the 117 EFL learners enrolled at the Academy. For this study, to prevent the potential effect(s) of participants’ previous

Table 1 Box design of the research

Pre-test		Treatment and Variables	Post-test 1		Treatment	Post-test 2
Testing Group	PPBT version	Administration Mode	CFLT1 with no item review option	Implementing the Equivalent Test after two weeks interval	Item Review	CFLT2 with item review option
		Age				
		Gender				
		ICT literacy				
		Computer Attitude				

onscreen test-taking experience on their subjective CFLT attitudes, only those with no personal experience of taking computerized tests were selected.

Consequently, of those 69 students who were originally identified, five were removed because they had experience of using a computerized test, and a further three others were unable to take part due to the place and time of the study. To measure the reading proficiency level of the participants, the remaining 61 students then underwent the Cambridge Reading Proficiency Test, as a result of which, three participants were excluded owing to a big difference between the ranges of their scores and those of the other test takers. The remaining 58 intermediate EFL learners were assigned to one testing group. Within the testing group, there was a higher distribution of males ($n=55\%$) as opposed to females ($n=45\%$) (Table 2).

The age of male and female participants ranged from 18 to 35 and 18 to 33, respectively. The mean age of male participants was 25.28 ($SD=4.84$) years, and that of females was 23.92 ($SD=5.59$) years. Consequently, the mean age of younger males (below-30) $M=22.45$ ($SD=2.63$) and older males (above-30) $M=31.50$ ($SD=1.43$) were calculated. For females, the mean ages of the younger participants (below-30) and older (above-30) were $M=20.50$ ($SD=2.28$) and $M=31.62$ ($SD=1.06$) years, respectively. When looking at the age profile of the testing group in its entirety, the mean ages were $M=21.57$ ($SD=2.63$) for younger participants (both genders) and $M=31.55$ ($SD=1.24$) for older ones.

3.3 Instruments

Since using inappropriate research data collection instruments can lead to collecting wrong and inappropriate data which could ultimately change the path of the research (Privitera, 2012), the researcher reviews the data collection instruments used in this research in respect of their validity and reliability values. The first of those instruments, the TOEFL general proficiency test was used to determine participants' language proficiency level and select homogenous participants. This test is considered to be a reliable and valid index of general English proficiency (PBT Complete Test/p.515–538) (Phillips, 2001). The test is composed of three sections including listening comprehension (35 min for 50 items), structure and written expression (25 min for 40 items) as well as vocabulary and reading comprehension (55 min for 50 items). The overall score of the test for each test taker was estimated by considering the total results for each module of the test including listening, structure, and vocabulary as well as reading comprehension. A scale ranged from 20–68 was used to report the obtained scores of each section. Then, the total score was reported based on the selected raw scores. TOEFL overall scores were reported on a scale that ranged from 217–677. Based on the Scoring Information (Phillips, 2001),

Table 2 Gender and age frequency distribution

		Frequency	Percentage			Frequency	Percentage	Total
Gender	Male	32	55.17	Female	26	44.83		58/100
Age	Younger	22	73.33	Younger	8	26.67		30/100
	Older	10	35.71	Older	18	64.29		28/100

the overall score of each section was determined through the converted score chart. The three obtained converted scores were added together to divide the received sum by 3. Then, the number was multiplied by 10 to attain the overall score for each participant. The EFL learners (117 EFL learners) were asked to complete the test in 115 min. Based on the general English language proficiency conversion table, the homogenous participants (the overall TOEFL score ranged from 450 to 510) were selected to participate in the main investigation. The descriptive statistics demonstrated that the total mean for the overall TOEFL score was equal to 485.74 ($SD=16.32$). Since the researcher planned to examine testing administration mode effect on reading performance of EFL learners, a more homogeneous group of participants in reading proficiency was needed. Accordingly, a separate reading comprehension test to assess reading skill proficiency was administered to the 61 test takers in order to explore their homogeneity in terms of their reading proficiency and exclude those participants with higher or lower difference in their reading comprehension performance. Then, to see if there was any difference between the mean of reading comprehension performance of the participants, their scores on Reading and Use of English Sample Test 1 from the Cambridge English Proficiency Sample Paper Tests package (CEP/SSU) (2015) was analyzed. The test composed of 53 questions from which the items from 1 to 24 were worth one point, 25 to 30 carried up to two points (two points were allocated to the questions 25–30 in this study), and two and one points were devoted to 31 to 43 and 44 to 53 question item sets, respectively. Test takers were supposed to finish the test in 90 min. The total scores (points) of test takers were calculated according to the 73 scores attained from the test. Based on the descriptive statistics, it was concluded that 61 participants had the same or approximate level of reading proficiency skill and subskills (based on the minimum score range of 44 and maximum 49) except three participants. Those who gained scores higher than 67 (Aria=67, Soroush=69, and AmirAli=71) were excluded from the study. After a one-week interval, the TOEFL paper-based reading comprehension pre-test from Phillips, D, 2001(p.343–349) (Phillips, 2001) composed of 50 question items was administered to the remaining 58 participants, with 55 min allocated for the test. Based on the results, no high dispersion of the scores from the mean score was observed for the participants.

Moreover, the mean score of $M=41.65$ indicated the same level of reading proficiency. The research data (scores) on participants' reading performance (TOEFL paper-based reading comprehension pre-test from Phillips, D, 2001(p.343–349)) was normally distributed due to Skewness and Kurtosis values (0.063 & -0.906) approximating to zero and Kolmogorov–Smirnov (0.200) as well as Shapiro–Wilk (0.508) significance values which were considered greater than 0.05 (Sig.>0.05). Additionally, a good internal consistency reliability $\alpha=0.853$ was reported by Cronbach's alpha coefficient.

The paper-based reading comprehension post-test from Phillips, D, 2001(p.452–460) (Phillips, 2001) was another research data collection instrument used in the research. Test takers should read each question and mark the right option on a separate answer sheet given to each test taker with the test papers. Each question item had only one correct answer and test takers needed to choose one option and mark the option on the answer sheet. If the test takers marked more than one option, the question was not scored, and the grade for that item was considered zero. If one of the two selected options by the test takers was the correct answer, it was not scored because more than

one option had been identified (this is in contrast with CFLT version of the test in which selection of only one option was possible automatically). The researcher scored the paper-based test papers. The test contained 50 multiple-choice questions and was administered to the participants in three different versions (Paper-Based/ Computer-Based version with item review option (CFLT1)/ Computer-Based version without item review option (CFLT2)) in three different testing sessions each with a four weeks interval to avoid the potential for practice effects.

Microsoft's word-processing division (Microsoft Word, 2010 version 201,004,220) was used to convert the PPBT version of the test into the CFLT version with an item review option (CFLT1). In this version, the passages and the multiple-choice questions were presented to the test takers on the screen, and they were able to navigate the passages and questions easily. They could scroll up and down through the whole test and check their answers. Test takers were required to read the questions appeared on the computer screen and choose the most appropriate option under each question by clicking the mouse on the blank space beside the options. Like the PPBT version of the test, test takers could review and change their answers by changing the tick from one selected option to another one. They could even go back to the previous pages to review and change their answers.

The PPBT version of the test was converted into the CFLT2 version using professional c# programming language. A Windows-based application was created using the c# programming language powered by Microsoft Visual Studio. In the CFLT2 environment, users could log in using their username and password. A demo allowed the test takers to learn how to use the platform. This was optional, however, and the test takers could skip it by clicking on the "Skip" button and go directly to the test. In the test itself, each passage was displayed on the left of the screen as a "Fixed" element so that test takers were able to read it while answering the related questions displayed on the right of the screen with a "Next Question" button below. By pressing the "Next Question" button below each question, the next question would be retrieved from the database, and there was no option for users to go back, review or change the items or their answers. Microsoft SQL server was used to save data in this application. At the end of the test, the test taker could see his/her total score on a result page by pressing the "Finish" button that was created by Crystal Reports, which was connected to the database. Although test takers could change their answers by clicking on the blank space beside the other options, after clicking on the "Next Question" button below the page, they could not go back and change their answer(s).

Consequently, they had no opportunity to review the items and modify their answers after going to the next question. Whereas in the PPBT version of the test, test takers marked the answers in a separate answer sheet by a pencil, in CFLT2 version of the test there was just one opportunity to mark an option as the answer. In all three versions, the first section of the test elicited biographical information such as the name, date, and place of the test, as well as the name of the test takers.

Computer attitude and its correlation with the testing performance were considered to be an independent variable in the research, and among several computer attitude scales, the Computer Attitude Scale (CAS) developed by Loyd and Gressard (Loyd & Gressard, 1985) as one of the most practical and popular research tools was

employed (Khoshsima et al., 2017b). This instrument collected data relating to attitudes towards the use of computer as a whole and was composed of 40 statements regarding four components including computer anxiety, computer confidence, computer liking, and computer usefulness (Loyd & Gressard, 1985). The general reliability coefficient of 0.95 for CAS calculated by Loyd and Gressard (1985) was the highest value among the other nine computer attitude scales (Hosseini et al., 2014). Then, reliability coefficients of 0.81, 0.86, 0.85 and 0.82 were attained for Computer anxiety, Computer Confidence, Computer Liking and usefulness (Woodrow, 1991), respectively, as the subscales of CAS. While the Scale was used by some researchers (Al-Amri, 2009; Hosseini and Hashemi Toroujeni, 2017; Stricker, Wilder and Rock, 2004; Yurdabakan & Uzunkavak, 2012), it was the first time that it was used in a Persian private EFL context; then, the internal consistency of the questionnaire was measured (Cronbach's alpha reliability coefficient, i.e., $\alpha = .711$).

The ICT Literacy Scale of TOEFL Examinees (henceforth LS) or TOEFL Familiarity Scale was a one-page questionnaire with 23 questions used to examine the correlation between ICT literacy and testing performance. Each question of the TOEFL questionnaire had four response options for which points from one to four were assigned. In this respect, a higher number attained from all 23 questions would indicate a greater degree of ICT literacy (Eignor et al., 1998). Examination of internal consistency of the scale resulted in $\alpha = .723$ Cronbach's alpha reliability coefficient.

3.4 Procedure

This quasi-controlled empirical study was conducted with a common-person and pretest–posttest design to examine the possible effect of the testing administration mode on Iranian intermediate EFL learners' reading comprehension performance in the Adrina Language Academy. In addition to the three versions of the TOEFL reading proficiency test to examine the differences in scores between the two modes and the effect of item review on the testing performance, two CAS (Computer Attitude Scale) and LS (ICT literacy Scale) of TOEFL Examinees questionnaires were used to examine the effect of a relatively wide range of variables on the comparability of paper-based and computerized assessments.

After implementing the English general proficiency test as the placement test and administering two phases of the reading proficiency test, 58 students were selected to participate in the study as the test takers. They were assigned to one testing group who were to take both versions of the same test in three versions (Paper-Based or PPBT/Computer-Based format with an item review option or CFLT1/Computer-Based format without an item review option or CFLT2) in three testing sessions with a four weeks interval after each testing session. The two four-week intervals between testing sessions were used to reduce the possible testing effects and the effect of testing information on the long-term memory of participants. Additionally, the impact of fatigue might also be mitigated. It is worth mentioning that the CAS and LS questionnaires were administered to the test takers to examine the correlation of computer attitude and ICT literacy with reading comprehension performance before taking the CFLT1. Some oral instructions were given to the test takers on how to take CFLT versions of the test at the

Table 3 Testing Sphericity Assumption

Measure: Reading Comprehension Performance							
Within-Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon Greenhouse–Geisser	Huynh–Feldt	Lower-bound
Testing Administration Mode	.990	.564	2	.754	.990	1.000	.500

beginning of the second (in addition to the instructive demo in the testing environment) and third testing sessions.

4 Results and discussion

Since the performance of the same participants on the dependent variable (reading comprehension scores) was measured three times over time, One-way within Subjects or Repeated Measures ANOVA was the main statistical test used for the current research design (Larson-Hall, 2010). First, the internal consistency for both PPBT and CFLT versions was calculated, and relatively high-reliability coefficients (PPBT, $\alpha=0.895$, CFLT1, $\alpha=0.883$ & CFLT2, $\alpha=0.923$) were achieved.

According to the Kolmogorov–Smirnov statistical test results, given $p=0.441$ for PPBT version, $p=0.489$ for CFLT with item review version (CFLT1), and $p=0.439$ for CFLT without item review version (CFLT2), it was concluded that each of the levels of the independent variables was normally distributed. Mauchly's test was run to see if **Sphericity was assumed (Sig. < 0.05)**. A statistically significance level of 0.05 was used to report all statistical analysis in the present research. According to the **Sphericity assumption**, equal variances should be received from the dependent variables of the related group(s) measured repeatedly. Mauchly's Test of Sphericity demonstrated that the Sphericity assumption was not violated, $\chi^2(2)=0.564$, $p=0.754$ (Table 3).

In three testing sessions, the mean score of CFLT1 ($M=43.26$, $SD=3.86$) was greater than PPBT ($M=41.49$, $SD=4.7$) by 1.77 and CFLT2 ($M=40.4$, $SD=4.65$) by 2.86 points (Table 4). On the other hand, the lower standard error of CFLT1 (0.507) indicated a relatively lower spread in the sampling distribution.

Table 4 Distribution scores in PPBT, CFLT1 & CFLT2 (Estimates)

Measure: Reading Comprehension Performance				
Testing Administration Mode	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
PPBT	41.492	.618	40.255	42.728
CFLT1	43.268	.507	42.252	44.284
CFLT2	40.404	.611	39.182	41.627

Table 5 RM ANOVA for finding an overall significant difference (Tests of Within-Subjects Effects)

Measure: Reading Comprehension Performance									
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power	
Testing Administration Mode	Sphericity Assumed	242.35	2	121.17	6.76	.002	13.53	.912	
	Greenhouse–Geisser	242.35	1.98	122.39	6.76	.002	13.40	.910	
	Huynh–Feldt	242.35	2.00	121.17	6.76	.002	13.53	.912	
	Lower-bound	242.35	1.00	242.35	6.76	.012	6.76	.725	
Error (Testing Administration Mode)	Sphericity Assumed	2041.07	114	17.90					
	Greenhouse–Geisser	2041.07	112.86	18.08					
	Huynh–Feldt	2041.07	114.00	17.90					
	Lower-bound	2041.07	57.00	35.80					

By examining the significance value obtained from Mauchly test, p -value=0.754 ($p > 0.05$), the equality of the observed variances of the differences between the existing levels were confirmed $\chi^2(2) = 0.564$, $p = 0.754$. Then, to answer the first part of the research question one, whether there is a statistically significant difference between the scores in three versions of the test (the effect of “Testing Administration Mode” on reading comprehension scores), the statistical results of One-Way RM-ANOVA were interpreted with a null hypothesis of no difference. Some have proposed that corrections such as Greenhouse–Geisser or Huynh–Feldt should be used even if the Sphericity assumption is met by Mauchly’s test (Howell, 2002) (whatever the results of Mauchly’s test). However, the researcher, in this study, have taken it for granted that RM-ANOVA is not as robust to the Sphericity violation. Based on the within-subjects effects table and “Sphericity Assumed” row output, the mean scores for reading comprehension scores across the three different PPBT, CFLT1, and CFLT2 versions were significantly different ($F(2, 114) = 6.76$, $p = 0.002$, < 0.05) (Table 5).

Since the results of Table 5 indicated a statistically significant difference in three mean scores, the Bonferroni post hoc test was run to find out where exactly the difference(s) happened. After finding out that the different testing administration modes did not have equal effects on reading comprehension performance and there was a statistically significant difference among three sets of scores (Sig=0.002, $P < 0.05$), the post-hoc test and Pairwise Comparisons results were used to find out which particular means exactly differed (Table 6).

From the results of Table 6, it was concluded that there was a significant difference in reading comprehension performance between CFLT1 and CFLT2 ($p = 0.002$). The results of RM-ANOVA Pairwise Comparisons indicated that the mean difference between just two CFLT1 ($n = 58$, $M = 43.26$, $SD = 3.86$) and CFLT2 ($n = 58$, $M = 40.40$, $SD = 4.65$) was statistically significant, Sig=0.002, $p < 0.05$. It was noted that, although CFLT1 differed from CFLT2, the mean difference between PPBT and CFLT1, and PPBT and CFLT2 was not statistically significant. It was concluded that no statistically significant difference was found between PPBT and CFLT1, and PPBT and CFLT2. This supports the Interchangeability or comparability of PPBT test scores and its computerized counterpart, and the first part of the first null hypothesis was confirmed.

Based on the descriptive statistics and **mean difference**, it can be seen that reading comprehension performance was significantly increased at CFLT1 compared to PPBT, but it surprisingly decreased at CFLT2 compared to CFLT1 as well as PPBT modes.

Table 6 Post-Hoc test (Pairwise Comparisons of three mean score differences)

Measure: Reading Comprehension Performance						
(I) Testing Administration Mode	(J) Testing Administration Mode	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
PPBT	CFLT1	-1.776	.752	.065	-3.632	.080
	CFLT2	1.087	.820	.571	-.936	3.111
CFLT2	CFLT1	-2.863*	.783	.002	-4.795	-.932

Table 7 Distribution of male/female test takers’ scores (Group Statistics)

	Gender	N	Mean	Std. Deviation	Std. Error Mean
PPBT	male	32	44.75	2.68	.474
	female	26	37.48	3.34	.655
CFLT1	male	32	43.76	3.87	.684
	female	26	42.65	3.84	.753
CFLT2	male	32	40.79	4.28	.758
	female	26	39.92	5.10	1.001

Based on the results, it would seem that the reason for the mean difference should be explored in more details in item review rather than testing administration mode factor.

To investigate the effect of gender and age moderator variables on the EFL learners’ reading comprehension performance in three testing versions, the results of Independent-Samples T-Test were used. Based on the results of descriptive statistics, male participants ($n=32$, $M=44.75$, $SD=2.68$) outperformed females ($n=26$, $M=37.48$, $SD=3.34$) in PPBT (Table 7). Accordingly, the results of the Independent-Samples T-Test indicated that there was a statistically significant difference in the reading comprehension scores for males and females in PPBT ($t(56)=9.19$, $p=0.000$). In CFLT1 with item review, male participants ($n=32$, $M=43.76$, $SD=3.87$) attained greater scores than female test takers ($n=26$, $M=42.65$, $SD=3.84$). This discrepancy in mean scores was compared by performing an Independent-Samples T-Test and surprisingly, no statistically significant difference was found between two groups’ performance on reading comprehension in CFLT1 with item review (CFLT1), $t(56)=1.08$, $p=0.283$. The same results were obtained for male ($n=32$, $M=40.79$, $SD=4.28$) and female ($n=26$, $M=39.92$, $SD=5.10$) participants in CFLT2 without any statistically significant difference between the mean scores $t(56)=0.69$, $p=0.488$.

Although the gender might be a factor to create performance difference on PPBT version, it could not be considered something that could have affected the reading comprehension performance of test takers in the CFLT1 session. This does not support the effect of gender on reading performance of test takers when they take

Table 8 Paired Samples Test to compare the mean difference between gender subgroups across PPBT & CFLT1

		Paired Differences				t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference			
					Lower	Upper		
Pair 1	Male PPBT vs Male CFLT1	.862	4.6	.813	-.796	2.521	1.061	31 .297
Pair 2	Female PPBT vs Female CFLT1	-5.158	5.249	1.029	-7.279	-3.038	-5.011	25 .000

Table 9 Distribution of younger/older test takers' scores (Group Statistics)

	Age	N	Mean	Std. Deviation	Std. Error Mean
PPBT	younger	40	40.51	4.87	.77
	older	18	43.65	3.54	.83
CFLT1	younger	40	42.62	3.92	.62
	older	18	44.69	3.40	.8
CFLT2	younger	40	40.46	4.55	.72
	older	18	40.27	4.97	1.17

computerized counterpart and the null hypothesis for the gender, the (a) section of the second part of the null hypothesis one, was confirmed.

From Table 8, it is also possible to conclude that there was not a statistically significant difference ($\text{Sig}=0.297$, $p>0.05$) between the reading performance of males on both PPBT and CFLT1 ($n=32$, $M=44.62$, $SD=2.64$ vs. $n=32$, $M=43.76$, $SD=3.87$). However, the Paired-Samples Test revealed that the difference between the females' scores received from PPBT ($n=26$, $M=37.50$, $SD=3.36$) and CFLT1 ($n=26$, $M=42.65$, $SD=3.84$) was statistically significant, $\text{Sig}=0.000$, $p<0.05$. Therefore, it can be concluded that females are more likely to enjoy the benefits of CFLT rather than PPBT.

The researcher considered the effect of gender on educational achievement within and beyond the group, especially reading performance when computerized testing was administered to EFL learners in a private context. It was seen that although male participants performed better in CFLT1 than female participants (within groups), their performance decreased in CFLT1 compared to PPBT (within groups). On the other hand, although female participants had lower performance on CFLT1 than the male participants in the testing group (beyond groups), they outperformed in CFLT1 compared to PPBT. From this, it was concluded that female participants might enjoy the benefits and advantages of CFLT rather than PPBT.

To investigate whether variations in age may create an equal or different performance in reading comprehension test, an Independent-Samples T-Test was used to compare the mean difference between younger (below-30) and older (above-30) age groups in PPBT, CFLT1, and CFLT2, separately. Table 9 provides descriptive statistics including mean and standard deviations in the reading comprehension performance for the two age groups of the participants in three testing sessions.

Independent-Samples T-Test results indicated a statistically significant effect of age on reading comprehension performance of the participants in PPBT session ($t(56)=-2.44$, $p=0.018$) with a better performance by the older age group ($n=18$, $M=43.65$, $SD=3.54$) than younger participants ($n=40$, $M=40.51$, $SD=4.87$) (beyond group). However, although the older age group performed better ($n=18$, $M=44.69$, $SD=3.40$) than the younger ($n=40$, $M=42.62$, $SD=3.92$) in the first CFLT in the second testing session, it was concluded that the effect of the age moderator variable was not statistically significant ($t(56)=-1.93$, $p=0.058$) (within groups). Finally, age was not found to be statistically significant in CFLT2 ($t(56)=0.13$, $p=0.891$), with the younger age group (below-30) outperforming insignificantly ($n=40$, $M=40.46$, $SD=4.55$) vs ($n=18$, $M=40.27$, $SD=4.97$) (within group). This

does not support the effect of age on reading performance of test takers when they take computerized counterpart, and the null hypothesis for the age, (b) section of the second part of the null hypothesis one, was confirmed. It might be said therefore that the younger and older groups had approximately the same performance in CFLT2 (Table 9); It can be said that age and gender do not have any effect on the computerized reading comprehension test with or without item review (especially CFLT1, due to its similarity in item review and the difference in the mode with PPBT).

In Table 9, the reading comprehension mean scores received from the three versions were carefully measured by 2×2 representation; age (older and younger participants)×gender (males and females) to verify the mean differences among the different age groups.

According to the distribution of younger and older participants’ scores on CFLT1, the mean score of younger participants on CFLT1 (M=42.62, SD=3.92) was higher than their mean score on PPBT (M=40.51, SD=4.84). Of the two PPBT and CFLT1 mean scores for younger participants, the highest mean score was found in CFLT1, with a relatively higher mean score by 2 points. On the other hand, the standard deviation in younger CFLT1 was higher than in PPBT; i.e., the dispersion of scores from the mean score of younger participants in CFLT1 was higher than in PPBT. It was therefore concluded that Standard Error of Measurement (SEM) in the CFLT1 by younger participants was lower than in PPBT (SEM/CFLT=0.62073 vs. PPBT=0.76628). Scores in the CFLTT1 were more consistent. Analysis of the scores of younger participants in PPBT and CFLT1 established the Sig. Observed value 0.034 at P<0.05. This level of significance value at 39 (N-1) degree of freedom in a 0.05 level revealed a statistically significant difference between the two sets of scores (Sig=0.034, P<0.05) (Table 10). In contrast, the older age subgroup performed better (n=18, M=44.69, SD=3.4) in CFLT1 than in PPBT (n=18, M=43.65, SD=3.54), and it was concluded that the main effect of the age moderator variable was statistically significant ($t(17)=-3.5, p=0.003$) (within groups). Based on Tables 9 and 10, the performance of both age groups was better in CFLT1, and it was concluded that both age groups might enjoy the benefits of computerized tests.

From the results, the greater mean scores of older male test takers (M=45.92, SD=3.27 & M=45.24, SD=2.73) revealed that this subgroup outperformed other subgroups in PPBT and CFLT1 versions of the reading comprehension test.

Table 10 Paired-Samples Test to compare the mean difference of age subgroups across PPBT & CFLT1

		Paired Differences				t	df	Sig.
		Mean	Std. Devia- tion	Std. Error Mean	95% Confidence Interval of the Difference			(2-tailed)
					Lower	Upper		
Pair 1	Younger PPBT vs Younger CFLT1	-2.12	6.1	.96	-4.07	-.17	-2.2 39	.034
Pair 2	Older PPBT vs OlderCFLT1	-18.41	22.15	5.22	-29.43	-7.39	-3.5 17	.003

Table 11 Testing Normality Assumption

One-Sample Kolmogorov–Smirnov Test		CAS Total Scale Score	TOEFL Familiarity Total Scale Score
N		58	58
Normal Parameters	Mean	92.0862	66.2069
	Std. Deviation	10.98370	7.98630
Most Extreme Differences	Absolute	.133	.084
	Positive	.103	.084
	Negative	-.133	-.074
Kolmogorov–Smirnov Z		1.013	.637
Asymp. Sig. (2-tailed)		.256	.812

Furthermore, although the standard deviation of the scores obtained from an older male subgroup in PPBT was higher than in the other subgroups, it was the lowest in CFLT1. The Pairwise Comparisons results of RM-ANOVA showed no statistically significant difference between PPBT and CFLT1 as well as CFLT2 (Table 6). Nevertheless, a statistically significant difference was found between two CFLT1 and CFLT2; the two versions that had the testing administration mode (computer) feature in common (but did not share the item review option). Since no statistically significant difference was found between the mean score in PPBT and both CFLTs with different modes, testing administration mode was not considered to be a variable that influenced participants' performance in the reading comprehension test and did not violate the reliability or validity of the test. It is worth mentioning that the two PPBT and CFLT1 versions had the item review option in common, and the better performance of the participants in CFLT1 with item review and the realized statistically significant mean difference between CFLT1 and CFLT2 might be attributed to the item review factor, given that both versions were in the computerized version. The results of the two male and female (gender) and the younger and older (age) groups' test takers on CFLT1 established a Sig. observed value of 0.283 and 0.058 at $P < 0.05$, respectively. This level of significance value at 56 (N-2) degree of freedom in a 0.05 level demonstrated that there was no statistically significant difference between the four sets of scores obtained from the CFLT1 version of the test. Accordingly, the male CFLT1 and female CFLT1 test scores, and younger CFLT1 and older CFLT1 scores were not different (Sig = 0.283 & 0.058, $P > 0.05$) in mean score. Although no statistically significant difference was found between the mean score of different age and gender groups in CFLT1, the older male subgroup performed better in CFLT1 in comparison to the other subgroups. On the other hand, conducted Paired-Samples T-Test results demonstrated that the mean score of males differs before taking CFLT2 (CFLT1 = $M = 43.76$, $SD = 3.87$) and after taking CFLT2 (CFLT2 = $M = 40.79$, $SD = 4.28$) at the 0.05 level of significance ($t(31) = 2.73$, $p = 0.010$). On average, their reading comprehension performance on CFLT1 was 2.97 points lower than after taking the CFLT2 version. There was also a significant decrease in the mean score of females before

($M=42.65$, $SD=3.84$) and after ($M=39.92$, $SD=5.10$) CFLT2; ($t(25)=2.38$, $p=0.025$). Based on the results, there was also a statistically significant difference in the mean scores for younger participants before ($M=42.62$, $SD=3.92$) and after ($M=40.46$, $SD=4.55$) taking CFLT2; ($t(39)=2.29$, $p=0.027$). A statistically significant difference was found for older participants in CFLT1 & CFLT2 ($t(17)=3.22$, $p=0.005$), although they had higher mean in CFLT1 ($M=44.69$, $SD=3.4$) vs ($M=40.27$, $SD=4.97$). There was strong evidence that all the participants' scores categorized into different subgroups based on age and gender were aggravated in CFLT2 compared to CFLT1. The aggravation or decrease in the reading comprehension scores on CFLT2 might be attributed to the item review option (not testing administration mode, age or gender difference), because testing administration mode, age and gender factors did not result in any change in the previous comparisons. This supports the effect of item review on reading performance of test takers when they take computerized counterpart, and the null hypothesis for the item review, (c) section of the second part of the null hypothesis one, was rejected.

Since ICT literacy and attitude towards using computer were considered in the current research as two moderator variables (predictors), their correlation with the CFLTs performance was investigated by adapting the ICT literacy Scale of (LS) TOEFL Examinees and the Computer Attitude Scale (CAS) questionnaires. The internal consistency of the Scale developed by Loyd and Gressard (Loyd & Gressard, 1985) with a high-reliability coefficient of 0.95 was measured, and the reported Cronbach's coefficient alpha 0.723 indicated good reliability for CAS (it consisted of 40 items) in the context of the current study. The Cronbach's alpha coefficient for the TOEFL ICT literacy Scale (LS) was an accepted value of 0.711 based on the acceptable ranges from 0.7 to 0.8 (Nunnally, 1978).

As the data collected from the two CAS and LS were transformed into two sets of total scale scores as continuous measures, Multiple Linear Regression was conducted to measure how much these predictor variables (PV) indeed affected the scores on CFLT1 independently. The p-value of Kolmogorov–Smirnov Goodness-of-Fit Test greater than the significance level indicated that the data followed a normal distribution (Table 11).

The graphical analysis was also used to check the normal distribution of the residuals to get a better insight into how far the variables deviated significantly from the normality assumption. Normal P-P Plot and histogram confirmed a fair normal distribution of the regression line residuals with no significant deviation (Fig. 1).

The presence of homoscedasticity was checked using scatterplot between residuals and dependent or response (RV) as well as predictor variables (PV). The scatterplots of related residuals to check the homoscedasticity assumption also indicated no tight or wide distribution on any sides of the plot. Then the approximately equal distribution of the points (neither clustered points nor especial pattern) on two sides of zero on both x and y-axes confirmed satisfaction of homoscedasticity assumption (Fig. 2).

Since the related residuals of the study were both normally distributed and homoscedastic, it was taken for granted that the PVs including ICT literacy and Computer Attitudes had the required pattern of straight-line correlation (Linearity) with the

Fig. 1 Normality Assumption Graphs

RV, i.e., CFLT1). However, the presence of a linear relationship between the response variable (CFLT1) and each of the predictor variables (CAS & LS), and between those collectively was checked by creating scatterplots. However, first, Pearson

Fig. 2 Homoscedasticity Assumption Graph

Table 12 Pearson Correlation Coefficients between variable pairs

	CFLT 1	CAS Total Scale Score	TOEFL Familiarity Total Scale Score
Pearson Correlation		.319	.150
Sig. (1-tailed)		.007	.130
N		58	58

Correlation was run, and the calculated coefficients demonstrated the amplitude of the correlation between RV and PV pairs. Based on the results, there was a statistically significant moderate positive linear correlation between attitudes toward the use of computers and CFLT1 reading comprehension scores, $r=0.319$. Moreover, the positive linear correlation between ICT literacy and CFLT1 was not statistically significant ($r=0.130$) (Table 12). Both R-values suggested moderate linear relationships between pairs of variables.

The visual inspection of the points distributed on the plot confirmed the presence of moderate linearity between the response (CFLT 1 scores) and predictor (CAS & LS) variables collectively (3D scatterplot) and separately (Fig. 3).

Then, according to the Test for Linearity output (ANOVA Table) (Table 13), Sig. Values of 0.762 & 0.713 > 0.05 were obtained as Deviation from Linearity for CAS and LS, respectively. Based on the results, the linear correlation between two pairs of CFLT1 and CAS, and CFLT1 and LS were confirmed.

Variance Inflation Factor (VIF) values ($VIF = 1.006 < 10$ (or less than 3) of both variables obtained from Coefficients table of our multiple linear regression model demonstrated that the predictor variables were not highly correlated with each other and Multicollinearity assumption was also satisfied (absence of multicollinearity). The VIF value of Collinearity Diagnostic statistics (one to ten) confirmed the absence of multicollinearity symptoms (similarity) between PVs in the model of the current research (Table 16). Moreover, a Durbin-Watson Test as a measure of autocorrelation presence over the related residuals probed the particular kind of serial correlation between the PVs, and based on the measurement; the 2.08 Durbin-Watson values confirmed that there was no autocorrelation (neither positive nor negative autocorrelation) between the variables (Table 14). Furthermore,

Fig. 3 Linearity Assumption Graph

Table 13 ANOVA output, Sig. for Deviation from Linearity (ANOVA Table)

			Sum of Squares	df	Mean Square	F	Sig
CFLT 1 CAS Total Scale Score	Between Groups	(Combined)	397.10	28	14.18	.932	.573
		Linearity	85.13	1	85.13	5.596	.025
		Deviation from Linearity	311.97	27	11.55	.760	.762
	Within Groups		441.16	29	15.21		
	Total		838.27	57			
			Sum of Squares	df	Mean Square	F	Sig
CFLT 1 TOEFL Famili- arity Total Scale Score (LS)	Between Groups	(Combined)	325.80	25	13.03	.814	.699
		Linearity	18.87	1	18.87	1.178	.286
		Deviation from Linearity	306.93	24	12.78	.799	.713
	Within Groups		512.46	32	16.01		
	Total		838.27	57			

heteroscedasticity (violation of homoscedasticity) as another part of MLR classical assumptions was tested in the model. The scatterplot graph was utilized to distinguish the size of the error term across the values of PVs (different or similar size of error term). Diffusion of the points on the plot showed no specific pattern

Fig. 4 Heteroscedasticity Assumption Graph

Table 14 A Model Summary table presenting R and R^2 values for CAS & LS (Model Summary)

Model ^b	R	R Square	Adjusted R Square	Std. The error of the Estimate	Durbin-Watson
1	.343 ^a	.118	.085	3.66746	2.084

^a Predictors: (Constant), TOEFL Familiarity Total Scale Score, CAS Total Scale Score

^b Dependent Variable: CFLT1

or shape, and consequently, the absence of the heteroscedasticity between variables was validated (Fig. 4).

According to the output, the researchers first examined if the model should be trimmed (i.e., insignificant predictor(s) elimination).

The R and R^2 values provided in the Model Summary table indicated the simple correlation value, $R=0.343$ (relatively good degree of correlation), and the total variation explanation of response variable (RV) (CFLT1/scores from the reading comprehension test) by predictor ones (CAS & LS), $R^2=0.118$. The R^2 value explained 11.8% variation in the RV caused by PVs (Table 14). The next table demonstrated a statistically significant prediction of RV done by the current regression model — the obtained Sig. Value $p=0.032$ ($p<0.05$) showed well that the regression model statistically significantly predicted the reading comprehension performance.

As it was mentioned, the achieved variance proportion of CFLT1 equal to 0.118 ($R^2=0.118$) could be described well by two CAS (computer attitude) & LS (ICT literacy) predictor variables (PVs). Technically, this value is the proportion of variation that is explained by the regression model. Then, Adjusted $R^2=0.085$ described that CAS and LS explained approximately 1% (0.085) of the CFLT1 variability (computerized reading comprehension performance). According to the categorizations proposed by Cohen (1988), this value (0.085/ approximately 1%) as the calculation of the effect size of PVs (CAS & LS) represented a low effect size of the PVs. Since it is common to report the research results based on R^2 , it was concluded that the utilized regression model was statistically significant in the current research. Then, $F(2, 55)=3.662$, $p=0.032<0.05$ demonstrated that the regression model statistically significantly predicted the RV (CFLT1). Due to the p -value (<0.05) for the F statistic, it was clear that at least one of the PVs is a statistically significant predictor of the RV, i.e., CFLT1 (Table 15).

Then, the Sig. Values provided in the Coefficients table helped to predict which PVs contributed statistically significantly to the model. The regression equation was

Table 15 The ANOVA table predicting RV (ANOVA)

Model ^a	Sum of Squares	df	Mean Square	F	Sig
Regression	98.509	2	49.255	3.662	.032 ^b
Residual	739.766	55	13.450		
Total	838.276	57			

^a Dependent Variable: CFLT1

^b Predictors: (Constant), TOEFL Familiarity Total Scale Score, CAS Total Scale Score

Table 16 The Coefficients table presenting which variable is significant (Coefficients)

Model ^a	Unstandardized Coefficients		Standardized Coefficients	t	Sig	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	29.20	5.543		5.269	.000		
CAS Total Scale Score	.108	.044	.309	2.433	.018	.994	1.006
TOEFL Familiarity Total Scale Score (LS)	.061	.061	.127	.997	.323	.994	1.006

^a Dependent Variable: CFLT1

shown based on the values received from the B column; $CFLT1 = 29.20 + 0.108$ (CAS) and $CFLT1 = 29.20 + 0.061$ (LS) (Table 16).

From Table 16, the participants' performance on the computerized reading comprehension test (CFLT1) could be predicted from just CAS. The Sig. Value (0.018) indicated that the attitudes of the participants towards the use of computer contributed statistically significantly to the model. To trim the regression model, LS (ICT literacy Scale assessed by TOEFL Familiarity Total Scale Score) was removed due to its insignificant contribution in the model, and because it was not found to be a significant predictor of the CFLT1 scores. The MLR (Multiple Linear Regression) determined that the test takers' familiarity with computers could not statistically significantly predict computerized reading comprehension performance (scores), $F(1, 56) = 3.662$, $p = 0.323 > 0.05$. The regression equation was as follows; predicted computerized reading comprehension scores ($CFLT1$) = $29.20 + 0.061$ (LS).

After removing LS as an insignificant predictor, the analysis was rerun considering just CAS as the PV. The output of the revised analysis showed $R^2 = 0.102$. The value described that the CAS as the PV explained 10.2% of the variation in the computerized reading comprehension scores as the DV. The R^2 value (0.102/10.2%) received from the regression model of Table 17 for CAS PV was approximately the same as the R^2 (0.118/11.8%) received from the results of preliminary model represented in Table 14, and consequently, it implied that the removed PV variable (ICT literacy /LS assessed by TOEFL ICT literacy or Familiarity Scale) was not effective in predicting the computerized reading comprehension scores (CFLT1 performance).

The value of $R = 0.319$ and the consequences of the R-square i.e. determination coefficient of 0.102 demonstrated that the computerized reading comprehension performance of the test takers (CFLT1) was affected by 10.2% by their attitudes towards the use of computers, while it was concluded that the part of this value

Table 17 Model Summary table presenting R and R^2 values for CAS (Model Summary)

Model ^b	R	R Square	Adjusted R Square	Std. The error of the Estimate	Durbin-Watson
1	.319 ^a	.102	.086	3.66728	1.932

^a Predictors: (Constant), CAS Total Scale Score

^b Dependent Variable: CFLT1

Table 18 The ANOVA table predicting RV by CAS (ANOVA)

Model ^a	Sum of Squares	df	Mean Square	F	Sig
Regression	85.135	1	85.135	6.330	.015 ^b
Residual	753.141	56	13.449		
Total	838.276	57			

^a Dependent Variable: CFLT1^b Predictors: (Constant), CAS Total Scale Score

(100%–10.2%=89.98%) could be attributed to the other variables, factors or causes (Table 17).

By forcing two PVs into MLR, the researchers removed LS as the insignificant variable and inserted the CAS data as the significant predictor into the Multiple Linear Regression in a rerun analysis. From the ANOVA output table (Table 18), the probability level of significance value of 0.015 was reported. This Sig. Value < 0.05 suggested that the current related MLR model predicted the reading comprehension performance of the participants on CFLT1. It was concluded that attitudes towards the use of the computer as a moderator variable in a computerized version of the test might influence the final performance of the test takers (Table 18).

The reported Sig. Value < 0.05 in the Coefficients table suggested that CAS was a significant predictor in the regression model (Table 19). From the table, the attitudes of test takers towards computers might have higher influence (Beta = 0.319) on a computerized reading comprehension test rather than ICT literacy whose non-statistically significant effect was confirmed in the preliminary model. Based on the significant value of $0.015 < 0.05$ for CAS predictor, it was therefore concluded that EFL learners' attitudes towards computers in a private context had a partially statistically significant effect on their performance on the computerized reading comprehension test (CFLT1). Accordingly, the researchers reached the firm conclusion that an increase or improvement or any positive change in the attitudes of the EFL learners towards the use of computers improve their performance on the computerized reading comprehension test.

The researchers examined the coefficients table to interpret the results. The prediction equation was based on the unstandardized coefficients as follows; $CFLT1 = 32.92 + 0.111 \text{ CAS}$ (Table 19). Moreover, while the constant value of 32.92 was observed as the predicted value of the response variable (CFLT1/Computerized Fixed-Length Linear Test of reading comprehension), the value

Table 19 The Coefficients table presenting if the predictor (PV) is significant (Coefficients)

Model ^a	Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig	Collinearity Statistics	
	B	Std. Error				Tolerance	VIF
(Constant)	32.926	4.101		8.029	0		
CAS Total Scale Score	0.111	0.044	0.319	2.516	0.015	1.000	1.000

^a Dependent Variable: CFLT1

of 0.015 was attained for CAS as the statistically significant predictor (predictor variable) of the study. This means that the predicted CFLT1 reading comprehension score for test takers with 0.015 attitudes towards the use of computer score was 32.92. Accordingly, the slope of CAS was 0.111 representing that any change or increase in the attitudes of the test takers towards the use of computers (let us say each unit increase in positive attitudes) would increase the predicted CFLT1 reading comprehension score by 0.111 units.

The results of multicollinearity assumption could be checked again for the CAS as the only significant predictor of the model. The tolerance value > 0.1 (1) or Variance Inflation Factor < 10 ($= 1$) indicated satisfaction of this classical assumption (Table 19). Also, the Normality Histogram was provided to check the normality of data received from CAS predictor (Fig. 5).

Finally, it was strongly concluded that the MLR model utilized in the current research indicated that EFL learners' attitudes towards the use of computers could statistically significantly predict their reading comprehension performance on the computer (CFLT1), $F(1,56) = 6.330$, $p = 0.015 < 0.05$ and attitudes towards computers accounted for 10.2% of the explained variability in CFLT1. The regression equation was considered as $CFLT1 = 32.92 + 0.111(\text{attitudes towards the use of computer assessed by CAS})$.

Based on the results, ICT literacy (assessed by LS) was not a statistically significant predictor of CFLT reading comprehension performance. The results

Fig. 5 Normality Assumption Graph

indicated that there was a statistically significant correlation between attitudes toward computer use and testing performance, and computer attitude was a statistically significant predictor of CFLT1 reading comprehension performance. Thus, based on the results, the null hypothesis for the ICT literacy was confirmed and the null hypothesis for the computer attitude moderator factor was rejected based on the evidence that computer attitude was a statistically significant predictor of the CFLT scores of intermediate EFL learners at the Adrina Language Academy.

5 Conclusion and implications

Before introducing computerized or onscreen version of any text, the performance in different modes should be compared and investigated. Performance (scores) of the EFL learners of ALA on a multiple-choice question type of reading comprehension skill received in three formats (PPBT, CFLT1 and CFLT2) have been analyzed to find out any statistically significant difference between two paper-based and computerized testing administration modes. PPBT and CFLT1 were administered to the participants to examine the effect of the “Mode” on the reading performance of the participants. The CFLT2 was administered to examine the effect of item review on their performance. Although some researchers have concluded that CFLT version of test resulted in lower scores than in paper-based tests (Chen et al., 2011; Clinton, 2019; Delgao et al., 2018; Golan et al., 2018; Lenhard et al., 2017; Rasmussen, 2015), in this research, the results of the participants’ testing performance in both PPBT and CFLT1 revealed that there was not any significant difference between the two sets of scores obtained from the two versions of the reading comprehension test that are in line with some studies (Farinosi et al., 2016; Hashemi Toroujeni et al., In Press; Hermena et al., 2017; Khoshshima et al., 2017h; Kong et al., 2018; Meyer et al., 2016; Priscari & Danielson, 2017). In fact, participants’ performances in two conventional paper-based and computerized testing sessions were not different. The findings are in contrast to the findings of some others who claim that they are not comparable (Aydemir et al., 2013; Delgao et al., 2018; Hosseini, et al., 2014; Khoshshima and Hashemi Toroujeni, 2017b; Pommerich, 2004; Singer & Alexander, 2017a, b). Singer and Alexander like Clinton (2019) state that their learners prefer reading texts in a print version and perform better in the paper-based reading comprehension test (Singer & Alexander, 2017a, b). In the current research, it was concluded that although PPBT and CFLT1 performances did not vary statistically, the participants absolutely performed better on CFLT1 due to their mean difference. Although the attitudes towards using computers was found to be an effective factor on the CFLT1 performance, the outperformance in the CFLT version of the test might also be attributed to the navigational issues such as rapid scrolling, mental reconstruction of knowledge and ideas expressed in the text or visual issues that were not controlled or investigated in the current research. These areas are therefore suggested for further investigation in the future. Whilst item review is a permanent feature of PPBT version, it might be missing for some CFLT models. Based on the findings, item review could have a significant effect on the performance of test takers. The performance of the participants in both PPBT and CFLT1 was better than their performance in CFLT2; in

fact, the CFLT2 scores reported the worst performance. Since no significant difference was found in either testing administration modes (PPBT & CFLT1), the better performance of the participants on PPBT and the significant difference between CFLT1 and CFLT2 might be attributed to the item review option. It was concluded that item review might result in an increase or improvement in the reading comprehension performance (PPBT/M=41.49, CFLT1/M=43.26, CFLT2/M=40.40). The findings of the current research are not in line with the findings of Revuelta that found no interaction between item review and testing performance (Revuelta et al., 2003).

Nowadays, it seems that younger students are more engaged in using technological tools and more attuned to their advantages. Their greater knowledge of digital mediums such as computers, tablets, and iPads may lead them to achieve better performance when reading onscreen. A further purpose behind the study was to probe the details of the difference in performance between male and female as well as younger and older participants who took both PPBT and CFLT1 versions of the test. Based on the findings, although the male group performed better than the female group in PPBT and a statistically significant difference was found between these two gender groups in the PPBT session, the findings demonstrated no significant difference in the reading comprehension performance of male and female participants in both CFLT sessions. It is worth mentioning that the male group performed better in both CFLT and PPBT. Generally speaking, because the purpose of the research was to examine the effect of gender on CFLT performance, finding no performance difference between the gender groups showed that this factor could not be associated with testing administration mode.

On the other hand, a comparison of the two sets of male scores on PPBT and CFLT1, and the two sets of female scores on PPBT and CFLT1 showed different results within the groups. Based on the findings, CFLT1 administration favored the female group rather than the male group. All the members of the female group performed better on the CFLT1 rather than in PPBT. However, the members of the male groups had approximately the same performance on both testing occasions and showed no significant difference in their performance across two paper and onscreen mediums (PPBT/M=44.62 vs. CFLT1/M=43.76, 1 point difference in their mean score). It could be concluded that the females might enjoy the benefits of the CFLT more than males. Moreover, changing testing administration mode may improve their performance (PPBT/M=37.5 vs. CFLT1/M=42.65). The first part of the findings of the current research on gender difference in the testing mode comparability study was compatible with the findings of Khoshshima, Hosseini and Hashemi Toroujeni in which gender was not found to be a factor that might have an impact on test takers' performance on CFLT (Khoshshima et al., 2017h). The CFLT1 performance difference between younger and older subgroups of participants (based on the definition of age) was also examined in the current research. Although the difference in performance between age subgroups in CFLT1 was fairly small, older participants had a higher mean score than the younger ones. Additionally, better performance of the older participants was confirmed with a statistically significant difference between two age subgroups. Surprisingly, the reading comprehension performance of both younger and older participants across CFLT1 was considerably better than their performance across PPBT. In particular,

the performance of older participants was found to be highly associated with the testing administration mode, and CFLT1 favored older participants much more than younger participants. Consequently, their performance was highly improved in CFLT1 (Older PPBT/ $M=26.27$ vs. Older CFLT1/ $M=44.69$). It was concluded that across CFLT1, older male participants performed better than the other age and gender subgroups ($M=45.24$), although the performance of this subgroup was rather negatively affected by CFLT1 (PPBT/ $M=45.92$ vs. CFLT1/ $M=45.24$). Younger male members were also slightly negatively affected by CFLT1. The male group with two age subgroups did not enjoy the performance on CFLT1. On the other hand, although younger and older female participants' reading comprehension performance on CFLT1 was lower than younger and older male participants (beyond groups comparison), their performance was absolutely positively affected by CFLT1 (within group comparison).

In the current research, ICT literacy was not stated as a first-rate contributor to differences in the performance between participants in CFLT1 reading comprehension. According to the findings of the research, ICT literacy was not seen as a reason behind the difference in the reading comprehension performance of the participants across two PPBT and CFLT1 modes. The findings are in accord with other similar studies such as Al-Amri's (Al-Amri, 2009) and Khoshshima and Hashemi Toroujeni's (Khoshshima & Hashemi Toroujeni, 2017) findings. Like this study, they found no statistically significant correlation or interactive effect between ICT literacy and testing performance. However, the impact of computer attitudes on the testing performance of the participants was confirmed. Technology is exerting a strong influence on education and assessment fields nowadays, conventional Paper/Pencil-Based Tests (PPBT) are being transformed into popular computerized version in many educational and testing contexts due to the several advantages it provides. In some educational centers, both versions are being used to satisfy educators' preference. The effect of testing administration mode (transforming from PPBT to CFLT) that is sometimes known as "Testing Mode Effect" or "Testing Administration Mode Effect" should be investigated to find out whether the existence of two versions is resulting in the same or equal results and scores. It is also the case that the effect of some external or internal moderator variables such as learning styles or strategies, computer anxiety, demographic features or attributes are worth investigating. In this research, no difference in the reading comprehension performance that happened across two PPBT and CFLT could be attributed to the mode effects. However, computer attitude (attitudes towards the use of a computer assessed by CAS) was distinguished as a moderator variable that could influence the CFLT performance of the participants. As other variables such as learning styles could not be addressed in this research, the researchers suggest examining how EFL learners with different learning styles or strategies and cognitive or metacognitive processes perform across different modes of testing. Finally, it is recommended that compute-adaptive language testing (CALT) as a subtest of computerized testing be considered as a reading comprehension skill domain. Unlike the common computerized tests, this flexi-level strategy provides the situations for the test-takers in which they answer just the

questions appropriate to their proficiency levels and the need to answer numerous difficult or easy questions is eliminated.

Funding This research was not supported by any specific grant from any public or private funding agency.

Code availability Microsoft's word-processing division (Microsoft Word, 2010 version 201.004,220) was used to convert the PPBT version of the test into the CFLT version with an item review option (CFLT1).

Declarations

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 the Persian consent form to sign. They were told to declare their agreement to participate in the study.

References

- ADB. (2018a). *Asian development outlook 2018: How technology affects jobs*. Asian Development Bank.
- ADB. (2017b). *Asian development outlook 2018: Transcending the middle-income challenge*. Asian Development Bank.
- Alakyleh, A. S. (2018). Evaluating the Comparability of (PPT) and (CBT) by Implementing the Compulsory Islamic Culture Course Test in the University of Jordan. *International Journal of Assessment Tools in Education*, 5(1), 176–186. <https://doi.org/10.21449/ijate.370494>
- Ala-Mutka, K. (2011). *Mapping digital Competence: Towards a conceptual understanding*. Seville: European Commission, Joint Research Centre. Institute for Prospective Technological Studies. Retrieved from <http://ipts.jrc.ec.europa.eu/publications/pub.cfm%3fid%bc4699>
- Al-Amri, S. (2009). *Computer-based testing vs. paper-based testing: Establishing the comparability of reading tests through the revolution of a new comparability model in a Saudi EFL context*. Thesis submitted for the degree of Doctor of Philosophy in Linguistics. University of Essex (UK).
- Al-Amri, S. (2007). *Computer-based vs. Paper-based Testing: Does the test administration mode matter*. Proceedings of the BAAL Conference.
- Alghammas, A. (2020). Online Language Assessment during the COVID-19 Pandemic: University Faculty Members' Perceptions and Practices. *Asian EFL journal*, 27(4.4): 169–195.
- American Educational Research Association. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.
- American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. American Psychological Association.
- Ary, D., Jacobs, L., Irvine, C., & Walker, D. (2018). *Introduction to research in education*. Cengage Learning.
- Aydemir, Z., Ozturk, E., & Horzum, M. B. (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.
- Backes, B., & Cowan, J. (2018). *Is the pen mightier than the keyboard? The effect of online testing on measured student achievement*. National Center for Analysis of Longitudinal Data in Education Research. Working Paper 190. April 2018.
- Bando, R., Gallego, F., Gertler, P., & Romero, D. (2016). Books or laptops? *The cost-effectiveness of shifting from printed to digital delivery of educational content*, No 22928, NBER Working Papers, National Bureau of Economic Research, Inc. Retrieved from <http://EconPapers.repec.org/RePEc:nbr:nberwo:2292>

- Bartram, D., & Hambleton, R. K. (2016). *The ITC guidelines: International standards and guidelines relating to tests and testing*. In F. T. L. Leong, D. Bartram, F. M. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC international handbook of testing and assessment* (p. 35–46). Oxford University Press. <https://doi.org/10.1093/med:psych/9780199356942.003.0004>
- Barzillai, M., & Thomson, J. M. (2018). Children learning to read in a digital world. *First Monday*, 23(10), 1–10.
- Bennett, S., Maton, K., & Kervin, L. (2008). The “Digital Natives” Debate: A Critical Review of the Evidence. *British Journal of Educational Technology*, 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>
- Blackhurst, A. E. (2005). Historical perspectives about technology applications for people with disabilities. In D. Edyburn, K. Higgins, & R. Boone (Eds.), *Handbook of special education technology research and practice*. (pp. 1–27). Knowledge by Design Inc.
- Blazer, C. (2010). *Computer-Based Assessments* (Vol. 0918). INFORMATION CAPSULE Research Services. Accessed online 30 December 2020. <https://files.eric.ed.gov/fulltext/ED544707.pdf>
- Boeve, A., Meijer, R., Albers, C., Beetsma, Y., & Bosker, R. (2015). Introducing computer-based testing in high-stakes exams in higher education: Results of a field experiment. *PLoS One*, 10(12), e0143616.
- Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices*. White Plains, NY: Pearson Education.
- Bugbee Jr. (1996). The Equivalence of Paper-and-Pencil and Computer-Based Testing. *Journal of Research on Computing in Education*, 28:3, 282–299. <https://doi.org/10.1080/08886504.1996.10782166>
- Burr, S. A., Chatterjee, A., Gibson, S., Coombes, L., & Wilkinson, S. (2016). *Key Points to Facilitate the Adoption of Computer-Based Assessments*. *Journal of Medical Education and Curricular Development*, 3, JMECD.S20379. <https://doi.org/10.4137/jme.cd.s20379>
- Carpenter, R., & Alloway, T. (2018). Computer versus paper-based testing: are they equivalent when it comes to working memory? *Journal of Psychoeducational Assessment*, 37(3), 382–394. <https://doi.org/10.1177/0734282918761496>
- Chen, D. & Catrambone, R. (2015). Paper vs. screen effects on reading comprehension, metacognition, and reader behavior. *Proceedings of The Human Factors and Ergonomics Society Annual Meeting*, 59(1), 332–336. <https://doi.org/10.1177/1541931215591069>
- 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>
- Chen, J., White, S., McCloskey, M., Soroui, J., & Chun, Y. (2011). Effects of computer versus paper administration of an adult functional writing assessment. *Assessing Writing*, 16, 49–71.
- Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology*, 33, 593–602. <https://doi.org/10.1111/1467-8535.00294>
- Clinton, V. (2019). Reading from paper compared to screens: A systematic review and meta-analysis. *Journal of Research in Reading*, 42(2), 288–325.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd ed.). Lawrence Earlbaum Associates.
- Csapo, B., Ainley, J., Bennett, R., Latour, T., & Law, N. (2012). “Technological issues of computer-based assessment of 21st century skills,” in *Assessment and Teaching of 21st Century Skills*, eds P. Griffin, B. McGaw, and E. Care (New York, NY: Springer), 143–230.
- Daghan, G. (2017). Views of Students about Technology, Effects of Technology on Daily Living and their Professional Preferences. *TOJET: The Turkish Online Journal of Educational Technology*, 16 (4).
- Delgado, P., Vargas, C., Ackerman, R., & Salmeron, L. (2018). Don’t throw away your printed books: A meta-analysis on the effects of reading media on reading comprehension. *Educational Research Review*, 25, 23–38.
- Dhawan, S. (2020). Online learning: A panacea in the time of COVID-19 crises. *Journal of Educational Technology*, 49(1), 5–22. <https://doi.org/10.1177/0047239520934018>

- Dogan, N., Kibrishoglu Uysal, N., Kelecioğlu, H., & Hambleton, R. K. (2020). An overview of e-assessment. *Hacettepe University Journal of Education*, 35(Special Issue), 1–5. <https://doi.org/10.16986/HUJE.2020063669>
- Doyle, M. (2020). WHO doctor says lockdowns should not be main coronavirus defense. <https://www.abc.net.au/news/2020-10-12/world-health-organization-coronavirus-lockdown-advice/12753688>. Accessed 26 April 2021.
- Eagly, A. H., & Shelly C. (1998). "Attitude Structure and Function." In Handbook of Social Psychology, ed. D.T. Gilbert, Susan T. Fisk, and G. Lindsey, 269–322. New York: McGowan-Hill.
- Eaves, R. C., & Smith, E. (1986). The effect of media and amount of microcomputer experience on examination scores. *Journal of Experimental Education*, 55, 23–26.
- 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).
- Eignor, D. R., Taylor, C., Kirsch, I., & Jamieson, J. (1998). *Development of a scale for assessing the level of computer familiarity of TOEFL examinees*. Educational Testing Service.
- Elliot, S. N., & Kettler, R. J. (2013). Test Accessibility: Item Reviews and Lessons Learned from Four State Assessments. *Education Research International*, 2013. <https://doi.org/10.1155/2013/952704>
- Emerson, L., & MacKay, B. (2011). A comparison between paper-based and online learning in higher education. *British Journal of Educational Technology*, 42(5), 727–735. <https://doi.org/10.1111/j.1467-8535.2010.01081.x>
- Farinosi, M., Lim, C., & Roll, J. (2016). Book or screen, pen or keyboard? A cross-cultural sociological analysis of writing and reading habits basing on Germany, Italy, and the UK. *Telematics and Informatics*, 33(2), 410–421.
- Florida Department of Education. (2006). What do we know about choosing to take a high-stakes test on a computer? Retrieved May 15, 2010, from: <http://www.fldoe.org/asp/k12memo/pdf/WhatDoWeKnowAboutChoosingToTakeAHighStakesTestOnAComputer.pdf>
- French, M. M. J., Blood, A., Bright, N. D., Futak, D., Grohmann, M. J., Hasthorpe, A., Heritage, J., Poland, R. L., Reece, S., & Tabor, J. (2013). Changing fonts in education: How the benefits vary with ability and dyslexia. *The Journal of Educational Research*, 106, 301–304.
- Galindo-Aldana, G., Meza-Kubo, V., Castillo-Medina, G., Ledesma-Amaya, I., Galarza-Del-Angel, J., Padilla-Lopez, A., and Moran, A. L. (2018) Computer-Based Neuropsychological Assessment: A Validation of Structured Examination of Executive Functions and Emotion. In: Harris D. (eds) *Engineering Psychology and Cognitive Ergonomics*. EPCE 2018. Lecture Notes in Computer Science, vol 10906. Springer, Cham. https://doi.org/10.1007/978-3-319-91122-9_26
- Gallagher, A., Bridgeman, B., & Cahalan, C. (2000). *The effect of computer-based tests on racial/ethnic, gender, and language groups* (GRE Board Professional Report No. 96–21P). Princeton, NJ: Education Testing Service.
- García Laborda, J., & Alcalde Penalver, E. (2018). Constraining issues in face-to-face and Internet-based language testing. *Journal for Educators, Teachers, and Trainers*, 9(2), 47–56.
- Garcia Laborda, J., Magal Royo, T., & Enríquez Carrasco, E. V. (2010). Teachers' trialing procedures for Computer Assisted Language Testing Implementation. *Eurasian Journal of Educational Research*, 39, 161–174.
- Gnamb, T. (2021). The development of gender differences in information and communication technology (ICT) literacy in middle adolescence. *Computers in Human Behavior*, 114(2021), 106533. <https://doi.org/10.1016/j.chb.2020.106533>
- Golan, D. D., Barzillai, M., & Katzir, T. (2018). The effect of presentation mode on children's reading preferences, performance, and self-evaluations. *Computers & Education*, 126, 346–358.
- Halamish, V., & Elbaz, E. (2019). Children's reading comprehension and metacomprehension on screen versus on paper. *Computers & Education*, 145. <https://doi.org/10.1016/j.compedu.2019.103737>
- Hancock, G., Schmidt-Daly, T., Fanfarelli, J., Wolfe, J., & Szalma, J. (2016). Is e-reader technology killing or kindling the reading experience? *Ergonomics in Design: the Quarterly of Human Factors Applications*, 24(1), 25–30. <https://doi.org/10.1177/1064804615611269>
- Hardcastle, J., Hermann-Abell, C. & DeBoer, G. (2017). *Comparing Student Performance on Paper-and-Pencil and Computer-Based-Test*. AAAS Project 2061 Science Assessment Website. Retrieved 30 December 2020, from <https://www.aaas.org/news/computer-based-vs-paper-and-pencil-tests-are-results-equivalent>

- Hashemi Toroujeni, S.M. (2016). *Computer-Based Language Testing versus Paper-and-Pencil Testing: Comparing Mode Effects of Two Versions of General English Vocabulary Test on Chabahar Maritime University ESP Students' Performance*. Unpublished thesis submitted for the degree of Master of Art in Teaching. Chabahar Marine and Maritime University (Iran).
- Hashemi Toroujeni, S.M., Thompson, N., & Faghihi, A. (In Press). Computer in Education Assessment: Exploring Score Equivalence of Paper-Based versus Computer-Based Language Testing considering Individual Characteristics. *Profile: Issues in Teachers' Professional Development*.
- Hensley, Kiersten Kenning. "Examining the effects of paper-based and computer-based modes of assessment on mathematics curriculum-based measurement." PhD (Doctor of Philosophy) thesis, University of Iowa, 2015. <https://doi.org/10.17077/etd.iireseh1q>
- Hermena, E. W., Sheen, M., AlJassmi, M., AlFalasi, K., AlMatroushi, M., & Jordan, T. R. (2017). Reading rate and comprehension for text presented on tablet and paper: Evidence from Arabic. *Frontiers in Psychology*, 8, 257.
- Horne, J. (2007). Gender differences in computerized and conventional educational tests. *Journal of Computer Assisted Learning*, 23, 44–55. <https://doi.org/10.1111/j.1365-2729.2007.00198.x>
- Hosseini, M. & Hashemi Toroujeni, S.M. (2017a). Validity and Reliability Examination in Onscreen Testing: Interchangeability of Scores in Conventional and Computerized Tests by Examining Two External Moderators. *International Journal of English Language Teaching and Linguistics (IJELTL)*, Vol 1(2), 2017. <http://ijeltal.com/index.php/ijeltal/article/view/17>
- Hosseini, M., Zainol Abidin, M. J., & Baghdarnia, M. (2014). Comparability of Test Results of Computer-Based Tests (CBT) and Paper and Pencil Tests (PPT) among English Language Learners in Iran. International Conference on Current Trends in ELT, 659-667. <https://doi.org/10.1016/j.sbspro.2014.03.465>
- Howell, D. C. (2002). *Statistical methods for psychology*. Pacific Grove, CA: Duxbury/Thompson Learning.
- Hughes, A. (2003). *Testing for language teachers*. (2nd ed.). Cambridge University Press.
- Jeong, H. (2012). A comparative study of scores on computer-based tests and paper-based tests. *Behaviour & Information Technology*, 33, 410–422.
- Jerrim, J. (2016). PISA 2012: how do results for the paper and computer tests compare? *Assessment in Education: Principles, Policy & Practice*, 23(4), 495–518.
- Jerrim, J., Micklewright, J., Heine, J. H., Salzer, C. H., & McKeown, C. (2018). PISA 2015: how big is the 'mode effect' and what has been done about it? *Oxford Review of Education*. <https://doi.org/10.1080/03054985.2018.1430025>
- Johnson, M., & Green, S. (2006). On-line mathematics assessment: The impact of mode on performance and question answering strategies. *The Journal of Technology, Learning, and Assessment*, 4(5).
- Karay, Y., Schaubert, S., Stosch, C., & Schüttelz-Brauns, K. (2015). Computer versus paper: Does it make any difference in test performance? *Teaching and Learning in Medicine*, 27(1), 57–62.
- Karim, M. R., & Hasan, M. (2020). Virtual Classes during COVID 19 Pandemic in Tertiary Level in Saudi Arabia: Challenges and Prospects from the Students' Perspective. *Asian EFL journal*, 27(5.1): 205–219.
- Katz, I. R., & Elliot, N. (2016). Information literacy in digital environments: Construct mediation, construct modeling, and validation processes. In *Information literacy: Research and collaboration across disciplines* (pp. 97–116).
- Khoshshima, H., & Hashemi Toroujeni, S. M. (2017). Transitioning to an Alternative Assessment: Computer-Based Testing and Key Factors related to Testing Mode. *European Journal of English Language Teaching*, 2(1), 54–74. <https://doi.org/10.5281/zenodo.268576>
- Khoshshima, H., & Hashemi Toroujeni, S.M. (2017b). Comparability of Computer-Based Testing and Paper-Based Testing: Testing Mode Effect, Testing Mode Order, Computer Attitudes and Testing Mode Preference. *International Journal of Computer (IJC)*, 24(1), 80–99. <http://ijcjournal.org/index.php/InternationalJournalOfComputer/article/view/825/4188>
- Khoshshima, H., & Hashemi Toroujeni, S.M. (2017h). Computer-Based Testing: Score Equivalence and Testing Administration Mode Preference in a Comparative Evaluation Study. *International Journal of Emerging Technologies in Learning (iJET)*, 12(10), 35–55. <https://doi.org/10.3991/ijet.v12i10.6875>
- Khoshshima, 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.

- Kılıçkaya, F. (2013). Computer-based grammar instruction in an EFL context: improving the effectiveness of teaching adverbial clauses. *Computer Assisted Language Learning*, 28(4), 325–340. <https://doi.org/10.1080/09588221.2013.818563>
- Kim, H. J., & Kim, J. (2013). Reading from an LCD monitor versus paper: Teenagers' reading performance. *International Journal of Research Studies in Educational Technology*, 2(1), 15–24. <https://doi.org/10.5861/ijrset.2012.170>
- Kong, Y., Seo, Y. S., & Zhai, L. (2018). Comparison of reading performance on screen and on paper: A meta-analysis. *Computers & Education*, 123, 138–149.
- Kumar, A. V. (2013). Automated Check-In and Scheduling System for a Web-Based Testing Application. All Graduate Plan B and other Reports. 288. <https://digitalcommons.usu.edu/gradreports/288>
- Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. Routledge.
- Leino, K. (2014). *The relationship between ICT use and reading literacy: focus on 15-yearold Finnish students in PISA studies (Academic dissertation)*. Jyväskylä University Press.
- Lenhard, W., Schroeders, U., & Lenhard, A. (2017). Equivalence of screen versus print reading comprehension depends on task complexity and proficiency. *Discourse Processes*, 54(5–6), 427–445. <https://doi.org/10.1080/0163853X.2017.1319653>
- Loyd, B. H., & Gressard, C. (1985). The Reliability and Validity of an Instrument for the Assessment of Computer Attitudes. *Educational and Psychological Measurement*, 45(4), 903–908. <https://doi.org/10.1177/0013164485454021>
- Mangen, A., & Kuiken, D. (2014). Lost in an iPad: Narrative engagement on paper and tablet. *Scientific Study of Literature*, 4(2), 150–177. <https://doi.org/10.1075/ssol.4.2.02man>
- Margolin, S. J., Driscoll, C., Toland, M. J., & Kegler, J. L. (2013). E-readers, computer screens, or paper: Does reading comprehension change across media platforms? *Applied Cognitive Psychology*, 27(4), 512–519.
- Meyer, A., Innes, S., Stomski, N., & Armson, A. (2016). Student performance on practical gross anatomy examinations is not affected by assessment modality. *Anatomical Sciences Education*, 9(2), 111–120.
- Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *ePIRLS 2016 international results in online informational reading*. Retrieved from Boston College. TIMSS & PIRLS International Study Center website <http://timssandpirls.bc.edu/pirls2016/international-results/>
- Newhouse, P. C., & Cooper, M. (2013). Computer-based oral exams in Italian language studies. *ReCALL*, 25(3), 321–339. <https://doi.org/10.1017/S0958344013000141>
- Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent?. *Ergonomics*, 51(9), 1352–1375 (p. 1362). <https://doi.org/10.1080/00140130802170387>
- Nunnally, J. C. (1978). *Psychometric theory*. (2nd ed.). McGraw-Hill.
- 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. Submission Date: 23/11/2017.
- Pagani, L., Argentin, G., Gui, M., & Stanca, L. (2016). The impact of digital skills on educational outcomes: Evidence from performance tests. *Educational Studies*, 42, 137–162. <https://doi.org/10.1080/03055698.2016.1148588>
- Phillips, D. (2001). *Longman Complete Course for the TOEFL Test: Preparation for the computer and paper tests*. Longman.
- 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.2012.1>
- Pieger, E., Mengelkamp, C., & Bannert, M. (2016). Metacognitive judgments and disfluency– Does disfluency lead to more accurate judgments, better control, and better performance? *Learning and Instruction*, 44, 31–40.
- Pokhrel, S., & Chhetri, R. (2021). A Literature Review on Impact of COVID-19 Pandemic on Teaching and Learning. *Higher Education for the Future*, 8(1), 133–141. <https://doi.org/10.1177/2347631120983481>
- 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).
- Pomplun, M., Frey, S., & Becker, D. F. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement*, 62(2), 337–354. <https://doi.org/10.1177/0013164402062002009>

- Porion, A., Aparicio, X., Megalakaki, O., Robert, A., & Baccino, T. (2016). The impact of paper-based versus computerized presentation on text comprehension and memorization. *Computers in Human Behavior*, 54, 569–579. <https://doi.org/10.1016/j.chb.2015.08.002>
- 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*. <https://doi.org/10.1016/j.chb.2017.07.044>
- Privitera, G. J. (2012). *Statistics for the Behavioral Sciences*. SAGE publication Inc.
- Purcell, K., Heaps, A., Buchanan, J., & Friedrich, L. (2013). *How teachers are using technology at home and in their classrooms*. Pew Research Center's Internet & American Life Project.
- Rasmusson, M. (2015). Reading paper–Reading screen: A comparison of reading literacy in two different modes. *Nordic Studies in Education*, 34(1), 3–19.
- 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>
- 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.
- 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).
- Reuelta, J., Ximénez, M. C., & Olea, J. (2003). Psychometric and psychological effects of item selection and review on computerized testing. *Educational and Psychological Measurement*, 63, 791–808.
- Sandford, A. (2020). Coronavirus: Half of humanity now on lockdown as 90 countries call for confinement. www.euronews.com, 03/04/2020, accessed 25 April 2021.
- Sangmeister, J. (2017). Commercial competence: Comparing test results of paper-and-pencil versus computerbased assessments. *Empirical Research in Vocational Education and Training*, 9(3). <https://doi.org/10.1186/s40461-017-0047-2>
- Sawada, Y. (2019). Infrastructure investments, technologies and jobs in Asia. *International Journal of Training Research*, 17(sup1), 12–25. <https://doi.org/10.1080/14480220.2019.1629724>
- Scheerens, J., Glas C., & Thomas, S. M. (2005). *Educational evaluation, assessment, and monitoring: A systemic approach*. Lisse: Swets & Zeitlinger B.V.
- Schroeders, U., & Wilhelm, O. (2011). Equivalence of Reading and Listening Comprehension across Test Media. *Educational and Psychological Measurement*, 74(5). <https://doi.org/10.1177/2F0013164410391468>
- Shraim, K. (2019). Online Examination Practices in Higher Education Institutions: Learners' Perspectives. *Turkish Online Journal of Distance Education*, 20(4), 185–196.
- Shute, V.J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning*, 33(1). <https://doi.org/10.1111/jcal.12172>
- Siddiq, F., Hatlevik, O. E., Olsen, R. V., Throndsen, I., & Scherer, R. (2016). Taking a future perspective by learning from the past – a systematic review of assessment instruments that aim to measure primary and secondary school students' ICT literacy. *Educational Research Review*, 19, 58–84. <https://doi.org/10.1016/j.edurev.2016.05.002>
- Siddiq, F., & Scherer, R. (2019). Is there a gender gap? A meta-analysis of the gender differences in students' ICT literacy. *Educational Research Review*, 27(2019), 205–217. <https://doi.org/10.1016/j.edurev.2019.03.007>
- Singer, L. M. & Alexander, P. A. (2017). 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, L. M., & Alexander, P. A. (2017). Reading across mediums: Effects of reading digital and print texts on comprehension and calibration. *The Journal of Experimental Education*, 85(1), 155–172.
- Singer Trakhman, L. M., Alexander, P. A., & Berkowitz, L. E. (2019). Effects of processing time on comprehension and calibration in print and digital mediums. *The Journal of Experimental Education*, 87(1), 101–115.
- Sintema, E. J. (2020, April 7). Effect of COVID-19 on the performance of grade 12 students: Implications for STEM education. *EURASIA Journal of Mathematics, Science and Technology Education*, 16(7). <https://doi.org/10.29333/ejmste/7893>
- Smolinsky, L., Marx, B. D., Olafsson, G., & Ma, Y. A. (2020). Computer-Based and Paper-and-Pencil Tests: A Study in Calculus for STEM Majors. *Journal of Educational Computing Research*, 58(7), 1256–1278. <https://doi.org/10.1177/0735633120930235>

- Stobart, G. (2012). Validity in formative assessment. In J. Gardner (Ed.), *Assessment and learning*. (pp. 233–242). Sage.
- Stole, H., Mangen, A., & Schwitter, K. (2020). Assessing children's reading comprehension on paper and screen: A mode-effect study. *Computers & Education*, 151(2020), 103861. <https://doi.org/10.1016/j.compedu.2020.103861>
- Stricker, L. J., Wilder, G. Z., & Rock, D. A. (2004). Attitudes about the computer-based Test of English as a Foreign Language. *Computers in Human Behavior*, 20, 37–54.
- Tahmasebi, S., & Rahimi, A. (2013). Computer-assisted assessment: Highlights and challenges. *Teaching English with Technology*, 13(2), 55–74.
- Tavakol, M., & Dennick, R. (2011). Making Sense of Cronbach's Alpha. *International Journal of Medical Education*, 2, 53–55.
- Terzis, V., & Economides, A. A. (2011). The acceptance and use of computer-based assessment. *Computers & Education*, 56(4), 1032–1044. <https://doi.org/10.1016/j.compedu.2010.11.017>
- Texas Education Agency. (2008). *A review of literature on the comparability of scores obtained from examinees on computer-based and paper-based tests*. Retrieved 28 December 2020, from <https://goo.gl/AAdc5o>
- Thurlow, M., Lazarus, S. S., Albus, D., & Hodgson, J. (2010). *Computer-based testing: Practices and considerations (Synthesis Report 78)*. University of Minnesota, National Center on Educational Outcomes.
- United Nations. (2020). *Policy brief: Education during COVID-19 and beyond*. United Nations. https://www.un.org/development/desa/dspd/wp-content/uploads/sites/22/2020/08/sg_policy_brief_covid-19_and_education_august_2020.pdf
- Vansickle, T. (2015). *Test Reliability Indicates More than Just Consistency*. QUESTAR ASSESSMENT, INC.
- Wang, S. D., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement*, 67(2), 219–238. <https://doi.org/10.1177/0013164406288166>
- Wang, H., & Shin, C. (2010). Comparability of computerized adaptive and paper-pencil tests. *Measurement and Research Service Bulletin*, 13, 1–7.
- Washburn, S., Herman, J., & Stewart, R. (2017). Evaluation of performance and perceptions of electronic vs. multiple-choice paper exams. *Advances in Physiology Education*, 41(4), 548–555. <https://doi.org/10.1152/advan.00138.2016>
- Woodrow, J. (1991). A Comparison of Four Computer Attitude Scales. *Journal of Educl. Computing Research*, 7, 165–187.
- World Health Organization. (2020). *Coronavirus*. Geneva.
- 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>
- Yurdabakan, I., & Uzunkavak, C. (2012). Primary school students' attitudes towards computer-based testing and assessment in turkey. *Turkish Online Journal of Distance Education*, 13(3), 177–188.
- Zhu, X., & Aryadoust, V. (2020). An investigation of mother tongue differential item functioning in a high-stakes computerized academic reading test. *Computer Assisted Language Learning*. <https://doi.org/10.1080/09588221.2019.1704788>

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

Authors and Affiliations

Seyyed Morteza Hashemi Toroujeni¹

¹ Konarak Department of Education, Iran Ministry of Education, Chabahar, Konarak, Iran