English Literature and Language Review
ISSN(e): 2412-1703, ISSN(p): 2413-8827
Vol. 3, No. 5, pp: 35-45, 2017
URL: http://arpgweb.com/?ic=journal&journal=9&info=aims

From Conventional to Technology-Assisted Alternative Assessment for Effective and Efficient Measurement: A Review of the Recent Trends in Comparability Studies

Monirosadat Hosseini Ph.D., Department of English Language, Faculty of Humanities Tehran, Payame Noor University, Tehran, Iran

Seyyed Morteza Hashemi Toroujeni* M.A. in TEFL, English Language Department, Faculty of Management and Humanities, Chabahar Marine and Maritime University, Chabahar, Iran

Abstract: There is no doubt that using computers in language testing as well as in language learning has some advantages and disadvantages. Despite the widespread use of computer-based testing, relatively few studies have been conducted on the equivalency of two test modes especially in academic contexts. However, some institutes and educational settings are going towards using computerized test due to its advantages without doing any comparability investigation beforehand. Perhaps because they mostly believe that if the items are identical, the testing mode is irrelevant. As the use of computerized test types is rapidly expanding, we need appropriate use of technology as a facet of language learning and testing. Regarding this accelerating development in computerized tests in language testing, further investigations are needed to ensure the validity and fairness of this administration mode in comparison with traditional one. This study provides a brief discussion on the importance of substituting CBT for PPT and necessity of doing comparability study before this transition. It presents the significance of the study followed by an illustration of the background of the comparability studies of CBT and PPT.

Keywords: Conventional testing; Computer-based testing; Comparability studies.

1. Introduction

One of the most appropriate ways of measuring students' learning in educational setting is assessment (Bachman, 2000). Portfolio assessment, performance assessment, self-assessment, and peer assessments are among the examples of different types of assessment (Peat and Franklin, 2002). In recent years, information and communication technology has been employed in assessment and examination to mechanize the testing process. Computer-Based Test (CBT) provides a variety of innovations in testing and can be used in different contexts; it is one of the important areas is language testing (Bennett, 1998). The history of computerized testing began in the early 1970s (Bachman, 2000; Bunderson *et al.*, 1989; Chapelle, 2007; Khoshsima *et al.*, 2017; Mazzeo and Harvey, 1988; Mead and Drasgow, 1993; Wainer *et al.*, 1990). With the appearance of new technologies, computerized testing has begun to be widespread and implemented in large scale testing (Higgins *et al.*, 2005; Khoshsima *et al.*, 2017). Examples include state drivers' license exams, military training exams, job application exams, entrance exams in postsecondary education, and certification exams by professional groups such as TOEFL or IELTS (Russo, 2002; Trotter, 2001). The limited accessibility to computer and high cost limited the implementation of computerized language testing in past years (Anakwe, 2008; Chapelle and Douglas, 2006; Paek, 2005) however, recent developments in communication technologies have created alternative test methods through computers and internet all around the world (Clariana and Wallace, 2002).

Computers has a critical role in our lives, the number of computer-delivered tests is increasing in language testing due to the perceived advantages of computer-delivered tests (Hashemi Toroujeni, 2016; Khoshsima *et al.*, 2017). Such developments in computer technologies have influenced many areas including educational settings such as online learning and testing (Bennett, 2002; Dooling, 2000; Pommerich, 2004). In language learning, the use of computers and electronic devices has become popular; especially in assessing the language proficiency of English learners, the most precise and available way is through computers (Bachman, 2000; Clariana and Wallace, 2002; Paek, 2005; Sawaki, 2001). However, the limited accessibility to computer and high cost of using computer in high stake tests had limited the implementation of computerized language testing (Clariana and Wallace, 2002).

Since 1990s, many researchers advocated the importance of assessment in helping students learn better (Earl, 2003; Hart, 1994; Leahy *et al.*, 2005; Marzano *et al.*, 1993; Popham, 2001; Wiggins, 1993). Earl (2003) describes

*Corresponding Author

examination in today‘s schools as primarily evaluation of learning. On the other hand, computerized testing advocators believe that traditional measurement implementation place too much emphasis on passing a test rather than on encouraging learners to learn beyond education (Tanner, 2001). However, as institutions started to accomplish computer-based testing in their examination systems, concerns arise about the comparability of scores from the two administration modes, PPT vs. CBT (Chapelle and Douglas, 2006; Clariana and Wallace, 2002; Wang, 2004). As the computerized tests have been using for almost 20 yearS, and the Computer Assisted Language Learning (CALL) has been common since the middle of 20th century, it has been necessary to develop the means to include computerized tests (Leahy *et al.*, 2005).

1.1. Computers and Language Testing

Although computer has played an important role in testing for more than 20 years, the literature on CALL has shown that there has been relatively little attention to Computer-Based Test (CBT) (Bachman, 2000; Sawaki, 2001). While computers have been important in language assessments, only a relatively small group of professional language testers uses computers in producing and validating language tests (Sawaki, 2001). Russell and Haney (2000) asserted that the ″mismatch between the mode of learning and assessment could cause achievements to be inaccurately estimated.″ (p.2).

As computers become increasingly available in educational settings, it is likely that English teachers will use them to administer tests (Trotter, 2001). Bennett (2002) believes that since computers entered in our lives and had integral role in education, and as developments in technology made measurement of constructs more possible, it is clear that the use of CBT for language testing will become increasingly inevitable in the future (Bennett, 2002). However, Norris (2000) raised the question about the comparability of CBT with PPT in language testing in that whether CBT can provide appropriate means to interpret the language skills or proficiencies tested according to language educators' purposes, and also whether it fulfils the intention of language testing uses.

Although CBT offers many advantages over traditional PPT (Khoshsima *et al.*, 2017; Poggio *et al.*, 2005; Russell and Haney, 1996; Sawaki, 2001; Zhang and Lau, 2006), assessment experts, researchers, practitioners, and educators have concerns about the equivalency of scores between the two test administration modes (Chapelle and Douglas, 2006; Douglas, 2000). To deal with this concern, many researchers conducted studies in synthesizing the administration mode effects on CBTs and PPTs (Clariana and Wallace, 2002; Hashemi Toroujeni, 2016; Higgins *et al.*, 2005; Johnson and Green, 2006; Khoshsima *et al.*, 2017; Olsen *et al.*, 1989; Paek, 2005; Poggio *et al.*, 2005; Pommerich, 2004; Salimi *et al.*, 2011; Zandvliet and Farragher, 1997; Zhang and Lau, 2006). Some researchers found that in comparability studies on CBT and PPT, test takers have done better on CBT (Bugbee and Brent, 1990; Clariana and Wallace, 2002; Lee *et al.*, 2010; Maguire *et al.*, 2010; Parshall and Kromery, 1993) and in some others test takers performed better on PPT (Al-Amri, 2008; Anakwe, 2008; Pomplun *et al.*, 2002; Salimi *et al.*, 2011).

1.2. Paper and Pencil Tests (PPT)

In this method, students often are assessed using paper and pencil. Not long ago, all tests were given using paper and pencil. Many teachers still choose to administer tests in this way. Teachers may tell you that there is little chance of having technical problems during the exam, students do not need to be familiar with computers, and there is no software to learn. However, for language test takers, who are more familiar with computer using in language learning, using paper and pencil test is considered less effective than computer based test where they need to be assessed in authentic situation more than other learners (Taylor A. R., 2005).

Some of the weaknesses of paper and pencil tests are the time it takes to grade and the ease with which students may copy or cheat. Rarely does a teacher make more than a few versions of a test for a given exam. Moreover, when the number of test takers is unlimited administering the exams, controlling the test takers and scoring them are more difficult (Paek, 2005). In this method of testing, the answer sheets of multiple-choice tests can be rated either by human raters including teachers, test developers or other stuffs or by mark reader machines. Then, the results of the tests, namely test scores, will be sent to students' files and teachers' record notebooks later. It may be done either through email, fax or other ways and usually there is a long span of the time between conducting the test and reporting the score.

1.3. Computer Based Test (CBT)

With the arrival of computer testing, there have been many changes in students' language assessment (or testing). Computer tests allow the teacher and test developers to design as many test versions as there are students and contents. With the use of computer rating, the tests can be immediately scored and transferred directly to the teacher‘s grade book and students’ file (Paek, 2005). This process can also eliminate some opportunities for human error. The teacher can determine which questions are missed, and more specifically which concepts are mastered or not mastered by using software (Bugbee, 1996). The results of some previous studies on the preference of students on CBT showed that they find it more credible, fair, interesting, promising, fast, and less difficult or stressful (Russell and Haney, 1996; Sambell *et al.*, 1999). Computerized tests can be conducted either over the internet or in a proctored setting (Choi and Tinkler, 2002).

2. Comparability Studies in Language Testing

Due to the variety of results of similar studies and the necessity of substituting CBT for PPT in some educational settings, especially in distance educational systems, where using electronic devices is inevitable, conducting comparability study is vital (Hashemi Toroujeni, 2016; Khoshsima *et al.*, 2017). Comparability of test scores should be examined before replacing or including CBT in the language assessment procedures (Pommerich, 2004). In addition, as computerized testing could be affected by the students' intention behaviour in using computer and their attitude and preference, doing some studies considering these variables in test results is important to see whether various testing modes examine the same construct without the interference of other construct irrelevant variables. For example, the literature on computer-assisted language learning indicates that both language learners and instructors have generally positive attitudes toward using computers in the classroom, but the evidence of their attitude towards the specific area of computer use, i.e. language testing is not enough (Esmaiel, 2006; Khoshsima *et al.*, 2017; Stricker and Attali, 2010). Reece and Gable (1982) found that identification of students with positive or negative attitudes towards the use of computers as well as methods that may influence those attitudes should be of great value for curriculum planners and program evaluators. Table 1 illustrates some differences between PPT and CBT drawn from the results of multiple studies in the comparison between the two test delivery modes. The table compares the similarities and differences in the process and appearance by test method. The interest here is to show how these differences in process and appearance may affect the results achieved. The table is divided into stages of the test process (distribution, administration, appearance/performance, responses, and the resulting data file).

Table-1. The summary of comparison of the test modality, PPT and CBT drawn from previous studies Test process

	Paper and Pencil Test	Computer-Based Test
Distribution	- printing/ addressing, shipping & mailing, errors may occur - Photocopying of forms is possible -Corruption of forms (e.g. trading between institutions, photocopy size and reflectivity, wrinkling, soiling) -Missing forms in the mail	- logon numbers and the website address sent by email to institution administrators - Logon numbers can be mixed up among grades or across institutions - Extra logon numbers are sometimes needed
Administration	- examiners are expected to distribute paper tests - A script is provided to the administrator - paper tests must be collected and returned	- institutions expected to distribute logon numbers, provide computer access, and direct students to the website - Internet or intranet access should be necessary - students must be brought to a computer lab to complete the test or equipped personally with specific facilities
Appearance/ Performance	- Single sheet of paper - Respondents fill in answer sheet with a pencil to indicate their response - Completion may be slower than computer	- Multiple screens - Respondents select responses by clicking the mouse over a radio button - Completion may be faster than paper - An indicator of progress through test administration is provided - a report of the test results can be provided to students after any exam
Responses	-No "rules" for data capture (e.g. respondents can fill in two circles, or mark half way between circles, etc.) - Corruption of forms (e.g. changing question wording or response scales) is possible - Notes on test papers (e.g. doodles, names, comments) are hard to capture or use -Existing the high possibility for cheating	- "Rules" for data capture are possible (e.g. no double response, no comments, no doodling etc.) - Corruption of forms not possible - Capture of comments can be accommodated for instructor's use - Additional District and questions can be added at the end - limiting the cheating chance by changing randomly the choices of any item computer by computer

Data File	- Scan to data file - the high possibility of mis-judging of students for their scores can occur - Hand entry required where form corruption prevents scanning (increase errors) - Existing much span of time in scoring and reporting to students	- Direct to data file (no need to scan) - Limited error introduced - not possibility of missing answer sheets - immediate scoring and immediate reporting to students - eliminating the possibility of mis-judging of students for their scores by showing the complete test report in computer screen

2.1. Validity of CBT

Fulcher (2009) states that, "Validity is conceptualized as test recognition by institutions, and validity evidence becomes the extent to which linkage is demonstrated by institutionally approved procedures (North *et al.*, 2003)".

The problem with CBT arises when the matter of validity comes; however, there is no evidence to show that the construct of CBT may produce less valid tests. Instead, there are other factors influencing a test that has little to do with the testing objectives the test developer had in mind (Chapelle and Douglas, 2006).

In addition, the validity of CBT is a crucial issue if we tend to employ it in testing certain language skills (Hosseini *et al.*, 2014). For example, in reading comprehension, CBT is at a disadvantage compared to many traditional paper-based tests because it can be boring to read from a computer screen due to the large parts of a text that cannot be viewed at one time, and browsing is complicated. Thus, it influences the comprehension results as the optimal energy and memory spend on just finding and following the whole text. In addition, each item in the test should measure the same proficiency or skill and subtests should test the various sub skills so that they can be distinguished from the main skill or proficiency. However, there is no guarantee that various sub skills are aspects of an overall skill if the items do not measure the same sub skill. This is why selecting a valid representation of sub skills must be taken into granted in language testing. Thus, tests that make use of item banks in CBT must have valid selection procedures (Fulcher, 2009).

2.2. Reliability of CBT

Reliability is the degree to which a test yields consistent and reproducible scores (Bachman and Palmer, 2000). According to the definition of reliability, CBT may have clear advantages over conventional methods of testing, of course, if the latter does not consist of objective scoring. Test takers in responding in pencil and paper tests may produce unpredictable results by marking their answer on an answer-sheet; in addition, mark reader machines may not be able to read certain or damaged sheets or in scoring by testing personnel some human errors may occur. In addition, test taker's responding in CBT may be much more controllable, because double answers or skipping the answers is prevented.

For past 15 years, language assessment researchers have been interested in computerization of L2 tests but few empirical studies have evaluated the comparability of the construct being measured in computerized and paper-based L2 tests and the generalizability of computerized test results to other conditions. Regarding this issue, Sawaki conducted a research in 2001 considering comparability of PPT and CBT of L2 reading test. The results showed that generalization of the findings to computerized L2 assessment was difficult. She emphasized the construct measured in such assessment contexts does not necessarily language skills, and mode of presentation studies in the non-assessment studies which involve L2 readers are scarce; in addition, there are limitations in the research methodologies used. The finding of her study suggests that examining the effect of test delivery mode and comparability studies should be an integral part of future construct validation of computerized tests of reading in a second/foreign language (Sawaki, 2001).

Therefore, one issue that requires prompt investigation is the effect of mode of presentation on comparability of the information obtained from computerized and paper-and-pencil tests. Chalhoub-Deville and Deville (1999), in their comprehensive study in analyzing the literature on comparability of CBT and PPT in L2 contexts, pointed out that there are scare studies in the comparability of PPT and CBT in L2 language tests. He also emphasized the importance of conducting comparability studies in local settings to detect any potential test-delivery-medium effect when a conventional test is converted to a computerized test. The requirement of interchangeability of test scores across modes of administration is one of the test score comparability objectives. Test score interchangeability does not imply that all test items necessarily perform the same in either mode. For example, some items might be easier or more difficult when administered on a computer. The summation of positive and negative effects over items may serve to counteract one another, resulting in a negligible effect on the total score.

2.3. Advantages of Computer-based Tests

Noyes and Garland (2008) believes that the benefits of standardized computer-based testing, such as quick and objective results as well as the ease of reporting results to others make this method very popular. Moreover, moves towards computerized testing is rooted from the advantages it provides in comparison with traditional paper-and-pencil format (Neumann and Baydoun, 1998; Pomplun and Custer, 2005; Salimi *et al.*, 2011; Terzis and

Economids, 2011; Yurdabakan and Uzunkavak, 2012). Such advantages, according to the findings of mentioned studies, include cost-effective administration, ease of administration, more accuracy, and immediacy of scoring and reporting, and flexible test scheduling. These studies, also, indicated that students who are familiar with computer feel more comfortable while using it (DeBell and Chapman, 2003; Higgins *et al.*, 2005; O'Malley *et al.*, 2005; Poggio *et al.*, 2005).

Because of its advantages, computerized testing now plays an important role in educational assessment (Clariana and Wallace, 2002; Hosseini *et al.*, 2014; Kingstone, 2009; Poggio *et al.*, 2005; Russell and Haney, 1996). It should be noted that some disadvantages are attributed to computerized testing as well, namely the higher costs of item development, which can outweigh many of the savings gained through the advantages (Noyes and Garland, 2008).

Some studies showed that students like to do computer-based test more than traditional method, because they find it more promising, fair, objective and credible (Al-Amri, 2009; Flowers *et al.*, 2011; Higgins *et al.*, 2005; Lightstone and Smith, 2009). Thus, studies are necessary to investigate the factors affecting the students' acceptance and intention to use CBT. The careful look at the discussions of some comparability studies showed that they emphasized on their suggestions about the need of more careful investigating the even weak differences in scores between CBT and PPT. They believed that some construct-irrelevant variables could influence the results of computerized test (Al-Amri, 2009; Bachman, 2000; Busch, 1995; Chapelle, 2007; Douglas, 2000; Stricker *et al.*, 2004) and should be considered important in similar studies.

3. Overview of Comparability Studies of CBT and PPT

Although computer-based testing has advantages over paper-based testing (Al-Amri, 2009; Clariana and Wallace, 2002; Khoshsima *et al.*, 2017), equivalency of two test modes should be ensured first (Clariana and Wallace, 2002; Paek, 2005; Sawaki, 2001; Wise and Plake, 1990). Due to the importance of this issue, the American Psychological Association (APA) assigned guidelines for CBT and its interpretations to retain the equivalency with PPT. Choi *et al.* (2003) defined equivalency as an investigation into the comparability of test modes or test tasks represented in different testing conditions. Neumann and Baydoun (1998) recommended equivalency of tests as: "the extent to which different formats of the same test measure the underlying trait determines whether they can replace each other.

Reviewing related literature on the comparability studies on CBT and PPT shows different results and opposite findings. In some of these studies, the test scores of two tests were similar (Anakwe, 2008; Bodmann and Robinson, 2004; Eid, 2005; Puhan *et al.*, 2007). Some others, in contrast, found different results with the priority of CBT over PPT and vice versa. For example, some studies that showed higher score on CBT such as Clariana and Wallace (2002). Contradictory findings reported lower performance on CBT than PPT (Choi and Tinkler, 2002; Flowers *et al.*, 2011; O'Malley *et al.*, 2005; Pomplun *et al.*, 2006).

Perhaps one of the most important reasons in the differences in test results in relation to the test mode effects of PPT and CBT is the difference in flexibility of test modes. Probably, it is because some CBTs do not provide the same level of flexibility as PPTs provide or vice versa. For example, some computer interfaces do not allow the student to skip, review, or change answers (Clariana and Wallace, 2002). Mason *et al.* (2001) also found evidence that shows the influence of different levels of flexibility on test results. There have been numerous works on the effect of changing answers on PPT results, and the results demonstrate that changing answers on multiple-choice tests in CBT slightly increases scores (Kruger *et al.*, 1977; Schwarz *et al.*, 1991; Vispoel, 1998).

However, some of these researchers attributed the differences to the similarity of the two test delivery modes. If computer-based tests closely are similar to paper-and-pencil format ones, the results could be similar as well. Evidence has accumulated to the point where it appears that in many traditional multiple-choice test settings, the computer may be used as a medium to administer tests without any significant effect on student performance (Paek, 2005). Any differences on multiple-choice tests, regarding the constructed response assessments, are related to individual basis. While most students prefer using the computer to paper format, their scores often vary depending on the mode of the test presentation (Choi and Tinkler, 2002; Parshall and Kromery, 1993). National Assessment of Education Progress (NAEP) in the Math online study suggests that performance on computer-based test items depends on the level of students' familiarity with using a computer. Students who are more familiar with the computer and more skilful in typing are more likely to perform better on the computer-based test. This finding suggests that computer familiarity may distort the measurement of mathematics achievement when tests are administered online to students who lack basic technology skills (Johnson and Green, 2006). This conclusion motivated researchers to examine the correlation between familiarity and test result on CBT while comparing two test modes.

There are several investigations regarding the comparability of test scores of computer-based tests and paper-based test among students in many fields of studies (Flowers *et al.*, 2011; Hargreaves *et al.*, 2004; Horkay *et al.*, 2005; Paek, 2005; Sandene *et al.*, 2005). Many focused on the differences between computerized tests and traditional paper and pencil tests without considering the effects of the learner adequately, if at all. One study by Fletcher and Collins (1986) listed the advantages and disadvantages of CBT over PPT without addressing the effects of learners' characteristics on test performance. On the other hand, there are few studies on the comparability of test modality of CBT and PPT in General English Language test performance in the form of multiple-choice form (Paek, 2005; Sawaki, 2001; Wallace and Clariana, 2000). However, some studies have been conducted on the relationship

of computer familiarity and attitudes of students with their test performance on CBT in their comparability studies of CBT and PPT (Al-Amri, 2008;2009; Boo, 1997; Clariana and Wallace, 2002; Tatira *et al.*, 2011; Taylor C. *et al.*, 1999; Yurdabakan and Uzunkavak, 2012; Zhang, 2007).

Clark (1983) according to the results of meta-cognitive studies on using technological devices in educational contexts found that students can achieve more benefits of their learning by using audio-visual or computer delivery instruments than conventional ones which is not generally related to the medium of instruction but the attitudes or strategies built into the learning materials. On the other hand, Hughes (2003) argues that the proper relationship between learning and testing is of that partnership. He also states that it is true that there may be situations where the teaching program is potentially good but assessment is not appropriate and vice versa. We are then, likely to suffer from harmful backwash from language testing when the learning occurs by technological devices such as computers in Computer Assisted Language Learning (CALL) environment but testing given traditionally in the form of paper and pencil test. Conversely, sometimes tests are taken through electronic devices but the instructions are traditionally provided to students during the course. Therefore, it is essential to do some comparability studies on CBT vs. PPT to deal with the issue of substituting computerized tests for traditional one to match the technological teaching with technological assessment in distance educational systems where using technology are is versatile and vital.

4. Research on Comparability of PPT and CBT

The concept of comparability is very important when test developers and researchers want to conduct studies on constructing and developing computer-based tests. Because despite the advantages of computer-based testing over paper-based testing, equivalency should be considered at first (Wise and Plake, 1990). Due to this, American Psychological Association (1986) proposed a set of instructions and guidelines to maintain the equivalency of paper and computer based test. When interpreting scores from the computerized versions of conventional tests, the equivalence of score from computerized version should be established and documented before using norms or cutting scores obtained from conventional tests. Scores from conventional and computer administration may be considered equivalent when (a) the rank orders of scores tested in alternative modes closely approximate each other, and (b) the means, dispersions, and shapes of the scores distributions are approximately the same by rescaling the scores from the computer mode (American Psychological Association (1986), cited in Sawaki (2001)). Russell and Haney (2000) conducted a study in the U.S. and indicated that students who are active computer users consistently perform on PPT. In their study, the students' responses written on computers were better than those written by hand. The findings were consistent across three subject areas, language arts, science, and math. When the students wrote on paper, only 30% performed at a `passing` level as compared to 67% of students who wrote on computers. In a second study, differences between computer-based and paper-based administrations measured a half standard deviation for students able to key at a minimum of 20 words per minute. Russell *et al.* (2003) suggested increased evidence that the tests that require students to produce written responses on paper underestimated the proficiency level of students who were accustomed to write at computers. The students performed better on authentic computer-based tests in other subject areas.

Similarly, Ricketts and Wilks (2002) three-year study introduced computer-based assessment of modules in biology to replace multiple-choice tests. The traditional multiple-choice test was given at the end of the first year, and then a computer-based online test was administered at the end of the second year. Students showed a 14% drop in scores on the computer-based assessment as compared to the results on the conventional multiple-choice test given the first year, even though the same questions were used. In the third year, performance on the computer-based assessment improved by 20% after implementing a question-by-question delivery. The course content remained the same, and the same lecturers delivered the material. The findings of the study showed that the results were not attributable to differences in the student cohort but rather to the student assessment interface. In contrast to the significant results obtained from the computer-based assessments in these two studies, the study by Hargreaves *et al.* (2004) yielded mixed results. The Hargreaves *et al.* (2004) study found that performance on mathematics computer-based assessments was better than on the paper and pencil tests. Each sample had a higher mean score on the computer test, 9.98 and 9.21, versus 9.03 on the paper administrations. However, the results were significantly better on one computer test. Two matched samples were assessed on one of the two mathematics pencil and paper tests and assessed one month later on one of the two computerized tests. The questions on the test were similar, but not identical. The study results did not indicate clearly any evidence of difference between using two different delivery test modes. In this study, the paper version was presented first; therefore, the results could have been influenced by carryover effect since the children had completed similar questions before completing the computer test one month later. The younger students did not appear to be affected by the unfamiliar presentation of an assessment on the computer (Hargreaves *et al.*, 2004).

Similarly, Bennett (2002) contended that for multiple-choice tests, the research to date has suggested that differences in computer experience have little, if any, effect on test scores. However, computer-based assessments incorporate more performance tasks, the complexity of responding could increase, and the digital divide may have an effect on scores. Disparities in academic achievement may increase because low-income and minority youth may unable to take full advantage of the educational benefits provided by computers and technology (Eamon, 2004).

Choi *et al.* (2003) defined comparability as "an investigation into the comparability of test methods or test tasks represented in different testing modes". Neumann and Baydoun (1998) also defined equivalency as "the extent to which different forms of the same test measure the underlying trait determines whether they can replace each other".

It means that if a test can measure the same construct, presenting in any form, PPT or CBT, then the different forms of the same test are equivalent and can be considered comparable. This is the reason that motivates the researcher to use these two words, 'comparability' and 'equivalency', interchangeably in the present study.

5. Conclusion

There have been studies on comparability of test results in PPT and CBT considering key factors associated with test results in different countries with different languages and technological backgrounds (Al-Amri, 2009; Bachman, 2000; Busch, 1995; Chapelle, 2007; Douglas, 2000; Flowers *et al.*, 2011; Yurdabakan and Uzunkavak, 2012). Nevertheless, there are disparities in the results of such studies. In addition, as far as the researcher knows, a few studies have been done recently in the context of universities in Iran. The researchers in the field have controversies whether test takers perform better on CBT or PPT. As it is necessary to refer to the stable results derived from comparability studies on transiting PPT to CBT and there is not yet one in the context of study, it is essential to conduct exclusive comparability studies in the language exams in Iran. In this regard, Salimi *et al.* (2011) have done a study on the importance of using computer-based test instead of paper-based test in testing English language in language institutes. According to their study, the results show that in spite of the positive attitude of students towards the use of computer, the respondents performed better on PPT. The other similar studies done by Khoshsima *et al.* (2017) were also a comparability study on CBT and PPT. Inevitably, with the widespread accessibility to computers in educational settings, universities in distance learning systems have focused the assessment on the use of computer-based testing (Fleming and Hiple, 2004). In this regard, Payame Noor University started to substitute CBT for PPT that caused concerns about the validity and comparability of scores from the two test modes. Administering computer-based exam in Payame Noor University since 2011 through System of Administering and Developing tests (SAD) system, yielded questions about the interchangeability of the results of two test modes while there has not been any investigation considering this issue.

The present study emphasized on the importance of substituting CBT for PPT and necessity of doing comparability study before this transition. It also reviewed the factors that may influence the test performance on CBT such as attitude, test mode preference, and familiarity with computers.

However, the results of numerous studies in the comparability of PPT and CBT show that there is no empirical evidence that identical paper-based and computer-based tests obtain the same results in all situations. The factors that influence the test results instead of the construct being measured are referred to as the ―test mode effect (Clariana and Wallace, 2002).

For example, paper-based test scores were greater than computer-based test scores for both mathematics and English tests in Mazzeo *et al.* (1991) study. While computer-based test scores were greater than paper-based test scores for a dental hygiene course unit midterm examination (DeAngelis, 2000) and some studies, in contrast, have reported non- significant difference between computer and paper-based tests (Boo, 1997; Mason *et al.*, 2001; Schaeffer *et al.*, 1993). The concern arises when it is asked: "How common is the test mode effective?"

References

Al-Amri, S. (2008). Computer-based testing vs. Paper-based testing: A comprehensive approach to examining the comparability of testing modes. *Essex Graduate Student Papers in Language & Linguistics*, 10: 22-44. Available: http://linguistlist.org/issues/19/19-2121.html

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).

American Psychological Association (1986). Guidelines for computer-based tests and interpretations. Author: Washington DC.

Anakwe, B. (2008). Comparison of student performance in paper-based versus Computer-based testing. *Journal of Education for Business*, 84(1): 13-17.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, 17(1): 1–42.

Bachman, L. F. and Palmer, A. S. (2000). *Language testing in practice*. 3rd edn: Oxford University Press: UK.

Bennett, R. E. (1998). *Reinventing assessment: Speculations on the future of large scale educational testing*. Educational Testing Service, Policy Information Center: Princeton, N. J.

Bennett, R. E. (2002). Inexorable and inevitable: the continuing story of technology and assessment. *The Journal of technology, Learning, and Assessment*, 1(1): 23.

Bodmann, S. M. and Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-based tests. *Jornal of Educational Computing Research*, 31(1): 51-60.

Boo, J. (1997). Computerized versus paper-and-pencil assessment of educational development: Score comparability and examinee preferences. Unpublished PhD dissertation, University of Iowa, USA.

Bugbee, A. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education*, 28(3): 282-300.

Bugbee, A. and Brent, F. (1990). Testing by computers: Findings in six years of use 1982-1988. *Journal of Research on Computing in Education*, 23(1): 87-100.

Bunderson, V., Inouye, D. and Olsen, J. (1989). *The four generations of computerized educational measurement. In R. L. Linn (Ed). Educational Measurement.* Oryx Press: Phoenix, AZ. 367-407.

Busch, T. (1995). Gender differences in self-efficacy and attitudes towards computers. *Journal of Educational Computing Research,* 12(2): 147-58.

Chalhoub-Deville, M. and Deville, C. (1999). Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics,* 19: 273-99. Available: https://www.cambridge.org/core/journals/annual-review-of-applied-linguistics/article/computer-adaptive-testing-in-second-language-contexts/928BA5479DAEF6146251EDA1AAD03FF9

Chapelle, C. (2007). Technology and second language acquisition. *Annual Review of Applied Linguistics,* 27: 98-114. Available: https://www.cambridge.org/core/journals/annual-review-of-applied-linguistics/article/technology-and-second-language-acquisition/0234CECFF2DDE14F8D1579567920B441

Chapelle, C. and Douglas, D. (2006). *Assessing language through computer technology.* Cambridge University Press: New York.

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

Choi, Kim, K. and Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing,* 20(3): 295-320.

Clariana, R. and Wallace, P. (2002). Paper-based versus computer-based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology,* 33(5): 593-602.

Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research,* 53(4): 445-59.

DeAngelis, S. (2000). Equivalency of computer-based and paper-and-pencil testing. *Journal of Allied Health,* 29(3): 161-64.

DeBell, M. and Chapman, C. (2003). *Computer and Internet use by children and adolescents in 2001: Statistical Analysis Report.* National Centre for Education Statistics: Washington, DC.

Dooling, J. (2000). What students want to learn about computers? *Educational Leadership,* 58(2): 20-24.

Douglas, D. (2000). *Assessing languages for specific purposes.* Cambridge University Press: Cambridge.

Eamon, M. K. (2004). Digital divide in computer access and use between poor and non-poor youth. *Journal of Society and Social Welfare,* 31(2): 91-112.

Earl, L. M. (2003). *Assessment as learning: Using classroom assessment to maximize student learning.* Corwin Press: Thousand Oaks, CA.

Eid, G. K. (2005). An investigation into the effects and factors influencing computer- based online math problem-solving in primary schools. *Journal of Educational Technology Systems,* 33(3): 223-40.

Esmaiel, Y. (2006). Theory in Practice: Constructivism and the technology of instruction in Authentic Project-Based computer Class. PhD thesis, The University of North Texas.

Fleming, S. and Hiple, D. (2004). Distance education to distributed learning: Multiple formats and technologies in language instruction. *CALICO Journal,* 22(1): 63-82.

Fletcher, P. and Collins, M. A. J. (1986). Computer-administered versus written test-advantages and disadvantages. *Journal of Computers in Mathematics and science Teaching,* 6: 38-43.

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

Fulcher, G. (2009). Test use and political philosophy. *Annual Review of Applied Linguistics,* 29: 3-20. Available: https://www.cambridge.org/core/journals/annual-review-of-applied-linguistics/article/test-use-and-political-philosophy/FCA93D9A5920AE096298B256B30E2CF6

Hargreaves, M., Shorrocks-Taylor, D., Swinnerton, B., Tait, K. and Threlfall, J. (2004). Computer or paper? that is the question: Does the medium in which assessment question are presented affect children‘s performance in mathematics? *Educational Research,* 46(1): 29-42.

Hart, D. (1994). *Authentic assessment: A handbook for educators.* Addison Wesley Publishing: Menlo Park, CA.

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 Arts in TEFL. Chabahar Marine and Maritime University (Iran).

Higgins, J., Russell, M. and Hoffmann, T. (2005). Examining the effect of computer-based passage presentation on reading test performance. *The Journal of Technology, Learning, and Assessment,* 3(4): 5-34.

Horkay, N., Bennett, R. E., Allen, N. and Kaplan, B. (2005). Online assessment in writing. In B. Sandene, N. Horkay, R. E. Bennett, N. Allen, J. Braswell, B. Kaplan, & A. Oranje (Eds.), online assessment in mathematics and writing: Reports from the NAEP Technology-Based Assessment Project (NCES 2005-457). Washington, DC: U.S. Department of Education, National Centre for Education Statistics.

Hosseini, M., Zainol Abidin, M. J. and 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. 659-67. Available: http://www.sciencedirect.com/science/article/pii/S1877042814025567?via%3Dihub

Hughes, A. (2003). *Testing for language teachers.* Cambridge University Press: Cambridge.
Johnson, M. and 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): 5-34.
Khoshsima, H., Hosseini, M. and Hashemi Toroujeni, S. M. (2017). Cross-mode comparability of computer-based testing (CBT) versus paper and pencil-based testing (PPT): An investigation of testing administration mode among iranian intermediate efl learners. *English Language Teaching,* 10(2): 23.
Kingstone, N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education,* 22(1): 22-37.
Kruger, J., Wirtz, D. and Miller, D. T. (1977). Counterfactual thinking and the first instinct fallacy. *Journal of Personality and Social Psychology,* 88(5): 725-35.
Leahy, S., Lyon, C., Thompson, M. and William, D. (2005). Classroom assessment minute by minute, day by day. *Educational Leadership,* 63(3): 19-24.
Lee, K. S., Osborne, R. E. and Carpenter, D. N. (2010). Testing accommodations for university students with AD/HD: computerized vs. paper-pencil/regular vs. extended time. *Journal of Education Computing Research,* 42(4): 443-58.
Lightstone, K. and Smith, S. M. (2009). Student choice between computer and traditional paper-and-pencil university tests: What predicts preference and performance? *International Journal of Technologies in Higher Education,* 6(1): 30-45.
Maguire, K. A., Smith, D. A., Brallier, S. A. and Palm, L. J. (2010). Computer-ased testing: A comparison of computer-based and paper-and-pencil assessment. *Academy of Educational Leadership,* 14(4): 117-25.
Marzano, R. J., Pickering, D. J. and McTighe, J. (1993). *Assessing student outcomes: Performance assessment using the dimensions of learning model.* Association for Supervision and Curriculums Development: Alexandria, V. A.
Mason, B. J., Patry, M. and Bernstein, D. J. (2001). An examination of the equivalence between non-adaptive computer-based and traditional testing. *Journal of Educational Computing Research,* 24(1): 29-39.
Mazzeo, J. and Harvey, L. A. (1988). *The equivalence of scores from automated and conventional education and psychological tests: a review of the literature. (Report No. CBR 87-8, ETS RR 88-21).* Educational Testing Services: Princeton, NJ.
Mazzeo, J., Druesne, B., Raffeld, P., Checketts, K. and Muhlstein, A. (1991). *Comparability of computer and paper-and-pencil scores for two CLEP general examinations. (College Board Report 91-5).* ETS: Princeton, NJ.
Mead, A. and Drasgow, F. (1993). Equivalence of computerised and paper-and- pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin,* 114(3): 449-58.
Neumann, G. and Baydoun, R. (1998). Computerization of paper-and-pencil tests: When are they equivalent? *Applied Psychological Measurement,* 22(1): 71-83.
Norris, J. M. (2000). Purposeful language assessment. *English Teaching Forum,* 38(1): 18-23.
North, B., Figueras, N., Takala, S., Van Avermaet, P. and Verhelst, N. (2003). Relating language examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF): Manual preliminary pilot version. Strasbourg, France: Language Policy Division of the Council of Europe.
Noyes, J. M. and Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics,* 51(9): 1352-75.
O'Malley, K. J., Kirkpatrick, R., Sherwood, W., Burdick, H. J., Hsieh, C. and Sanford, E. E. (2005). Comparability of a paper based and computer based reading test in early elementary grades. Paper presented at the AERA Division D Graduate Student Seminar, Montreal, Canada.
Olsen, J., Maynes, D., Slawson, D. and Ho, K. (1989). Comparison of paper- administered, computer-administered and computerized achievement tests. *Journal of Educational Computing Research,* 5(3): 311-26.
Paek, P. (2005). Recent trends in comparability studies. Pearson Educational Measurement Research Reports. Research Report 05-05. Pearson Educational Measurement. USA.
Parshall, C. G. and Kromery, J. D. (1993). Computer versus paper and pencil testing: An analysis of examinee characteristics associated with mode effect. Abstract from: ERIC Abstract No. ED363272. paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.
Peat, M. and Franklin, S. (2002). Supporting student learning: the use of computer based formative assessment modules. *British Journal of Educational Technology,* 33(5): 515-23.
Poggio, J., Glasnapp, D., Yang, X. and Poggio, A. (2005). A comparative evaluation of score results from computerised and paper & pencil mathematics testing in a large scale state assessment program. *The Journal of Technology, Learning and Assessment,* 3(6): 5-30.
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): 4-43.
Pomplun, M. and Custer, M. (2005). The score comparability of computerized and paper-and-pencil formats for K-3 reading tests. *Journal of Educational Computing Research,* 32(2): 153-66.
Pomplun, M., Frey, S. and Becker, D. (2002). The score equivalence of paper- and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement,* 62(2): 337-54.
Pomplun, M., Ritchie, T. and Custer, M. (2006). Factors in paper-and-pencil and computer reading score differences at the primary grades. *Educational Assessment,* 11(2): 127-43.

Popham, W. J. (2001). *The truth about testing: An educator's call to action.* Association for Supervision and Curriculum Development: Alexandria, VA.

Puhan, G., Boughton, K. and Kim, S. (2007). Examining differences in examinee performance in paper and pencil and computerised testing. *The Journal of Technology, Learning, and Assessment,* 6(3): 5-20.

Reece, M. J. and Gable, R. K. (1982). The development and validation of a measure of general attitudes toward computers. *Educational and Psychological Measurement,* 42(3): 913-16.

Ricketts, C. and Wilks, S. J. (2002). Improving student performance through computer-based assessment: Insights from recent research. *Assessment & Evaluation in Higher Education,* 27(5): 475-79.

Russell, M. and Haney, W. (1996). Testing writing on computers: Results of a pilot study to compare student writing test performance via computer or via paper-and-pencil. Retrieved July 12, 2011, from ERIC database.

Russell, M. and Haney, W. (2000). The gap between testing and technology in schools. The National Board on Education Testing and Public Policy. Retrieved from ERIC database.

Russell, M., Goldberg, A. and O'Conner, K. (2003). Computer-based testing and validity: a look back into the future. *Assessment in Education,* 10(3): 279-93.

Russo, A. (2002). Mixing Technology and Testing, Computer-Based Testing, The School administrator. Available: http://www.aasa.org/SchoolAdministratorArticle.aspx?id=10354,9.05.2012

Salimi, H., Rashidy, A., Salimi, A. H. and Amini Farsani, M. (2011). Digitized and non-digitized language assessment: A comparative study of Iranian EFL Language Learners. International Conference on Languages, Literature and Linguistics IPEDR 26.

Sambell, K., Sambell, A. and Sexton, G. (1999). *Student perceptions of the learning benefits of computer-assisted assessment: a case study in electronic engineering. In S. Brown, P. Race, & J. Bull (Eds.), Computer assisted assessment in higher education.* Kogan Page: London.

Sandene, B., Horkay, N., Bennett, R. E., Allen, N. B., J., Kaplan, B. and Oranje, A. (2005). *Online assessment in mathematics and writing: Reports from the NAEP technology based assessment project (NCES 2005-457).* U.S. Department of Education, National Center for Education Statistics: Washington, DC.

Sawaki, Y. (2001). Comparability of conventional and computerized tests of reading in a second language. *Language Learning & Technology,* 5(2): 38-59.

Schaeffer, G., Reese, C., Steffen, M. M., R. and Mills, C. (1993). *Field test of a computer-Based GRE general test. Reports-Research/Technical ETS-RR-93-07.* Educational Testing Services: Princeton.

Schwarz, S. P., McMorris, R. F. and DeMers, L. P. (1991). Reasons for changing answers: An evaluation using personal interviews. *Journal of Educational Measurement,* 28(2): 163-71.

Stricker, L. J. and Attali, Y. (2010). Test Takers' Attitudes about the TOEFL iBT™. TOEFL iBT Research Report, TOEFLiBT-13. Available: http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2010.tb02209.x/abstract

Stricker, L. J., Wilder, G. Z. and Rock, D. A. (2004). Attitudes about the computer-based test of english as a foreign language. *Computers in Human Behavior,* 20(1): 37-54.

Tanner, D. E. (2001). Authentic assessment: A solution, or part of the problem. *High School Journal,* 85(1): 6-13.

Tatira, B., Mutambara, L. H. N., Chagwiza, C. J. and Nyaumwe, L. J. (2011). Computerized summative assessment of multiple-choice questions: Exploring possibilities with the zimbabwe school examination council grade 7 assessments. *Computer and Information Science,* 4(6): 66-74.

Taylor, A. R. (2005). A future in the process of arrival: Using computer echnologies for the assessment of student learning: Society for the Advancement of Excellence in Education.

Taylor, C., Kirsch, I., Eignor, D. and Jamieson, J. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning,* 49(2): 219-74.

Terzis, V. and Economids, A. A. (2011). Computer based assessment: Gender differences in perceptions and acceptance. *Computers in Human Behavior,* 27(2011): 2108-22.

Trotter, A. (2001). Testing firms see future market in online assessment. *Education Week on the Web,* 20(4): 6.

Vispoel, W. P. (1998). Reviewing and changing answers on computer-adaptive andself-adaptive vocabulary tests. *Journal of Educational Measurement,* 35(4): 328-45.

Wainer, H., Doran, N., Flaugher, R., Green, B., Mislevy, R., Steinberg, L. and Thissen, D. (1990). *Computer adaptive testing: A primer.* Lawrence Erlbaum Associates: Hillsdale. NJ.

Wallace, P. E. and Clariana, R. B. (2000). Achievement predictors for a computer-applications module delivered via the world-wide web. *Journal of Information Systems Education,* 11(1): 13-18.

Wang, S. (2004). *Online or paper: Does delivery affect results? Administration mode comparability study for stanford diagnostic reading and mathematics tests.* Harcourt Assessment Inc: USA.

Wiggins, G. P. (1993). *Assessing student performance: Exploring the purpose and limits of testing.* Jossey-Bass Inc: San Francisco.

Wise, S. and Plake, B. (1990). Computer-based testing in higher education. *Measurement and Evaluation in Counselling and Development,* 23(1): 3-10.

Yurdabakan, I. and Uzunkavak, C. (2012). Primary school students' attitudes towards computer based testing and assessment in Turkey. *Turkish Online Journal of Distance Education,* 13(3): 177-88.

Zandvliet, D. and Farragher, P. (1997). A comparison of computer-administered and written tests. *Journal of Research on Computing in Education,* 29(4): 423-38.

Zhang (2007). EFL Teachers' Attitudes toward Information and Communication Technologies and Attributing Factors. Peking University.

Zhang and Lau, C. A. (2006). A comparison study of testing mode using multiple-choice and constructed-response items – Lessons learned from a pilot study. Paper presented at the Annual Meeting of the American Educational Association, San Francisco, CA.