Available online at www.jlls.org

JOURNAL OF LANGUAGE AND LINGUISTIC STUDIES

ISSN: 1305-578X

Journal of Language and Linguistic Studies, 14(1), 67-85; 2018

Computer-based and paper-based testing: Does the test administration mode influence the reliability and validity of achievement tests?

Hüseyin Öz [a] *, Tuba Özturan [b]

[a] *Department of Foreign Language Education, Hacettepe University, Ankara 06800, Turkey*
[b] *School of Foreign Languages, Erzincan University, Erzincan 24000, Turkey*

APA Citation:

Öz, H., & Özturan, T. (2018). Computer-based and paper-based testing: Does the test administration mode influence the reliability and validity of achievement tests? *Journal of Language and Linguistic Studies, 14*(1), 67-85.

Submission Date: 23/11/2017

Acceptance Date: 06/03/2018

Abstract

This article reports the findings of a study that sought to investigate whether computer-based vs. paper-based test-delivery mode has an impact on the reliability and validity of an achievement test for a pedagogical content knowledge course in an English teacher education program. A total of 97 university students enrolled in the English as a foreign language (EFL) teacher education program were randomly assigned to the experimental group that took the computer-based achievement test online and the control group that took the same test in paper-and-pencil based format. Results of Spearman Rank order and Mann-Whitney U tests indicated that test-delivery mode did not have any impact on the reliability and validity of the tests administered in either way. Findings also demonstrated that there was not any significant difference in test scores between participants who took the computer-based test and those who took the paper-based test. Findings were discussed in terms of the idea that computer technology could be integrated into the curriculum not only for instructional practices but also for assessment purposes.

Keywords: Computer-based testing; paper-based testing; reliability; validity; English teacher education

1. Introduction

With the introduction of the digital revolution, educators have begun to benefit from modern computer technology to carry out accurate and efficient assessment of learning outcomes both in primary/secondary and higher education. In recent years, Turkish institutions of higher education have also started integrating e-learning and assessment initiatives into their undergraduate programs. It is assumed that Turkish educational institutions will gradually move components of their assessment systems to online delivery or computerized mode. There are several reasons for implementing computerized assessments in education. We can reduce the "lag time" in reporting scores, increase the efficiency of assessment, achieve the flexibility in terms of time and place, give immediate feedback and announce students' scores immediately, analyze student performance that cannot be investigated from paper-based tests by implementing individualized assessments customized to student needs and

* Corresponding author. Tel.: +90-312-780-5521
E-mail address: hoz@hacettepe.edu.tr
This article is part of the second author's master thesis, completed with the supervision of the first author.

minimize the paper consumption and cost as well as duplicate or mail test materials (Alderson, 2000; Bennett, 2003; Noyes & Garland, 2008; Paek, 2005; Roever, 2001). This paper reports on findings of a study that investigated whether computer-based and paper-based tests as test delivery modes would influence the reliability and validity of the achievement test for a pedagogical content knowledge course in an English as a foreign language (EFL) teacher education program.

1.1. Reliability and validity criteria of tests

Defining the aims of tests and choosing the most suitable test type should be done before administering a test. However, these are not enough in order to have an effective test. In this sense, educators have to first consider some specific principles. Validity and reliability are foremost among these principles. As the most essential criterion for the quality of any assessment, validity is the relation between the aim and the form of the assessment and refers to whether a test truly measures what we claim it to measure. In other words, the tests measure what they are supposed to measure once the tests are valid (Fulcher & Davidson, 2007; Stobart, 2012). As it is a very crucial criterion for conducting tests, this following question lingers: how can instructors create valid tests or increase the validity of tests? There are some tips available to them, documented in available academic literature. Firstly, direct testing should be done whenever feasible, and explanations should be made clear. Secondly, scoring should be directly in relation to the targets of tests. Lastly, reliability has to be satisfied. Otherwise, validity cannot be assured (Hughes, 2003).

Reliability, on the other hand, is the degree to which a test measures a skill and/or knowledge consistently (Scheerens, Glas, & Thomas, 2005, p. 93). Therefore, similar scores are commonly achieved on a reliable test once the same exam is administered on two different days or on two different but parallel formats. It is important to note that Brown and Abeywickrama (2010) and Hughes (2003) both emphasize that the interval between the administrations of two tests should be neither too long as students might learn new things nor too short as it might change students' ability to remember the exam questions. Once the test is reliable, the test-takers will get more or less the same score no matter when the test is administered, on a certain day or on coming days, and teachers have to prepare and administer reliable tests so as to obtain similar results from the same students, but at a different time (Hughes, 2003, p. 36). Reliable tests give predictions about to what extent measurement-related factors may have impact on test scores. These factors can be grouped into the following categories: test factors that refer to the clarity of instruction, items, paper-layout and the length of the test; situational factors that refer to the conditions of the room; and individual factors that cover the physical and psychological state of test-taker. All these factors should be considered while interpreting the reliability of any test scores.

1.2. Computer-based testing alternatives

Computers are undoubtedly part of our daily lives; they take part in many different walks of life actively. This role change in computer applications goes back to the late 1970s. Since then, computers have had a vital place in the world, especially for educational purposes. In addition to the widespread use of web and computers as teaching sources both inside and outside the class (especially for distance education), computers have come to offer testing alternatives for teachers as well. Today, it is estimated that nearly 1000 computer-assisted assessments are done each day in the UK (Lilley, Barker, & Britton, 2004). These assessment models do not only refer to the traditional tests that are administered on computers in class under the supervision of proctors. It has different sorts of alternatives which are named as computer-based testing (CBT), web-based testing (WBT) and computer-adaptive testing (CAT). These are briefly introduced below.

Computer-based testing roughly refers to making use of computers while preparing questions, administering exams and scoring them (Chapelle, 2001), and with the advent of using computers as testing devices since the 1980s, a different point of view has been gained so that more authentic, cost-saving, scrutinized and controlled testing environment can be achieved, comparing to traditional paper-and-pencil based one (Jeong, 2014; Linden, 2002; Parshall, Spray, Kalohn, & Davey, 2002; Wang, 2010; Ward, 2002). Computer-based testing, which started in the late 1970s or in the early 1980s, was always thought as an alternative to paper-based testing (Folk & Smith, 2002), because "one size fits all" solution across testing programs was not desired at all (Ward, 2002, p. 37).

Computers have brought many advantages. First of all, they have the potential to offer realistic test items like media, graphics, pictures, video and sound (Chapelle & Douglas, 2006, p. 9; Linden, 2002, p. 9). Therefore, students can be involved in a real-life testing environment where there are many integrated activities. In other words, students can respond to computers orally, draw on the screen while answering the question, see and interpret graphics or tables for an open-ended question and so on, and handicapped test-takers can take the exams on computer with great ease. CBT also supplies immediate feedback and scoring (Chapelle & Douglas, 2006; Parshall et al., 2002), which has significant impact over pedagogy (test-takers can grasp their mistakes when immediate feedback is offered upon the completion of the test) and eases teachers' workload of scoring all papers – teachers may spend much time on scoring exam papers, and also, generally they cannot give enough feedback about each student's mistakes, or even if they provide feedback, it may be so late that students do not remember the questions or their answers. Another issue that should be mentioned here is that especially for open-ended questions, subjective-scoring may be in due. However, thanks to computer technology, objective scoring can be achieved, and problems caused by handwriting disappear, too. And the last important feature of CBT or Computer-Assisted Assessment (henceforth CAA) is that the examiners can collect data about the exam such as how many questions have been answered correctly, how many of them have been omitted and how many minutes have been spent for each question, which is called as *response latency* (Parshall et al., 2002, p. 2).

Since the beginning of using computers as testing tools, many different computer-based test delivery modes have come to scene: computer-adaptive testing (CAT), linear-on-the-fly testing (LOFT) or computerized fixed tests (CFT), computerized mastery testing (CMT) (Ward, 2002, p. 38) and automated test assembly (ATA) (Parshall et al., 2002, p. 11). CAT is totally performance or individual based testing. The more a candidate answers questions correctly, the more challenging questions appear on the screen, and vice versa. On the contrary, LOFT or CFT has fixed time and test-length for all test-takers. Exam security is the main goal in LOFT, rather than having psychometric values as in CAT (Parshall et al., 2002). As for CMT, it aims to divide test-takers into mastery and non-mastery groups (Folk and Smith, 2002, pp. 49-50). Lastly, ATA chooses items from an item pool in regard to the test plan and makes use of content and statistical knowledge. This kind of test has fixed time and is not in adaptive mode (Parshall et al., 2002, p. 10).

Kearsley (1996) emphasized the importance of web and its future potential as an educational tool many years ago. Not only is Web a means of delivering information, material, news and so on from one part of the world to the whole, but also it is the most commonly used and significant benefit of teachers for a variety of things like searching different types of materials, teaching for distance education, presenting, preparing tests and delivering them. The reason lying behind this change is that since 1990s, international connectivity has not been limited only to teaching staff at universities and to their use of network in computer labs, and without any doubt, it has brought many differences. As for the testing applications, universal access to computer-assisted assessment has been introduced, and a bulk of opportunities for autonomous learning and self-assessment has spread all around the world,

and so have computer-based applications. Today, thanks to web-based applications, students and teachers can be universal and universally in touch (Chapelle, 2001, p. 23).

As an alternative of CAA, web-based testing is specifically driven and delivered by means of web, and it means that the tests can be taken anywhere and anytime, which constitutes the great advantage over traditional paper-based and computer-based tests (Roever, 2001). Moreover, the web system also makes it possible to create unique exams, and it is based on an important mathematical content (McGough, Mortensen, Johnson, & Fadali, 2001). As Roever (2001, pp. 90-91) mentions that WBT is threefold as low-stakes assessment, medium-stakes assessment and high-stakes assessment, which can address for different needs: low-stakes tests are used to give feedback about examinees' performances over a certain subject or skill. The examinees can take these tests wherever they want. On the other hand, medium-stakes assessment covers midterm and final exams done in classes, placement tests or any tests that have impact on the examinees' lives. These kinds of tests are carried out by proctors in a lab. And lastly, high-stakes assessment is the one the results of which may affect greatly the examinee's life like being accepted to a university or certification programs or citizenship tests and so on. Among these three types, WBT is much more useful when it is done for low-stakes assessment.

In three phases (preparation, delivery and assessment), a question can be created on the web. Accordingly, an item is on the threshold of being created at *authoring time.* Teachers can prepare questions and store them in an item bank by using web tools. Then, questions or items are selected in order to conduct the test. The selection of the items is done either statistically by teachers themselves or dynamically by the system at run time (Brusikolovsky & Miller; 1999, p. 2). After delivering the items and conducting the exam, examinees' answers are assessed as correct, incorrect or partially correct. On the web technology, preparing, delivering and assessing questions are based on HTML codes (Brusikolovsky & Miller, 1999, pp. 2-3).

The last mode of CAA, computer-adaptive testing (CAT) that is based on each student's performance during the exam has been utilized for many years. The cycle of CAT begins with a question that is neither so easy nor so difficult. According to the answer of each test-taker to the item, which question to be asked from the item pool is decided. More clearly, if a test-taker answers a question correctly, the next one will be harder or on equal difficulty. On the contrary, if a test-taker answers a question incorrectly, the next one will be easier. Hence, CAT is said to be based on performance (Chapelle, 2001; Flaugher, 2000; Guzmand & Conejo, 2005; Lilley et al., 2004), and definitely, this new individualized exam model (Wainer & Eignor, 2000, p. 1) offers more confidential testing atmosphere for both teachers and students (Guzman & Conejo, 2005; Linden & Glas, 2002). Students can see each item on screen at a time, and they cannot skip the questions. While the test-takers are busy with each question, the system calculates the scores and decides which question will be next in relation to the previous answers given by the test-takers (Brown, 2004; Hughes, 2003). This measurement model in CATs is known as Item Response Theory (IRT) or Latent Trait Theory, the mathematical bases of which were outlined by Lord and Novick around the 1970s (Stevenson and Gross, 1991, p. 224; Tung, 1986, pp. 4-5).

The idea lying behind IRT goes back to the psychological measurement model, put forward by Alfred Binet and today known as the Stanford-Binet IQ test (Linden & Glas, 2002). Binet's idea of measuring each test-taker separately and according to their performance while they are taking the test has been accepted as the only adaptive testing approach for more than fifty years (Cisar, Radosav, Markoski, Pinter, & Cisar, 2010), but there was one drawback stated about this smart system: despite its truly adaptive side, experienced and skilled teachers (examiners) might be needed in order to administer large-scale tests. Therefore, it was practical only for small-scale tests (Madsen, 1991). Today, CAT is used not only for small-scale exams but also for large-scale high-stakes exams as well. For example, Graduate Management Admission Test, Microsoft Certified Professional and Test of

English as a Foreign Language have been administered in the CAT mode (Lilley et al, 2004, p. 110), and SIETTE is a web-based CAT system used in Spain (Guzman & Conejo, 2005, p. 688).

Many schools and universities have started to benefit from web technology while administering exams. One of them is Iowa State University that has created the WebCT. This smart system does not require any technical information so as to use it, and teachers can easily create and publish online courses and exams (Chapelle & Douglas, 2006, p. 63). Among other online tools to be utilized are Hot Potatoes, Discovery School Quiz Center, Blackboard and Questionmark (Chapelle & Douglas, 2006, pp. 72-73).

1.3. Studies on comparability of reliability and validity by test mode

Over the last two decades a number of comparability studies have concentrated on the effects of the test delivery mode on student performance, i.e., whether the test scores obtained from computer- and paper-based tests are interchangeable; these are referred to as "mode effects" (Bennett, 2003; Choi, Kim, & Boo, 2003; Dunkel, 1991; Paek, 2005; TEA, 2008; Wang, Jiao, Young, Brooks, & Olson, 2007). These studies often revealed mixed results regarding the comparability issues of CBT and PBT in different content areas. Some studies show that CBTs are more challenging than PBTs (Creed, Dennis, & Newstead, 1987; Laborda, 2010) or vice versa (Chin,1990; Dillon, 1994; Yağcı, Ekiz & Gelbal, 2011), whereas some studies conclude that CBTs and PBTs are comparable (Akdemir & Oğuz, 2008; APA, 1986; Bugbee, 1996; Choi, *et al.*, 2003; Choi & Tinkler, 2002, cited in Wang & Shin, 2009; Higgings, Russell, & Hoffmann, 2005; Jeong, 2014; Kim & Hyunh, 2007; Logan, 2015; Muter, Latremouille, Treurniet, & Beam, 1982; Paek, 2005; Parshall & Kromrey, 1993; Retnawati, 2015; Russell, Goldberg, & O'conner, 2003; Stevenson & Gross, 1991; Tsai & Shin, 2012; Wang et al., 2007; Wang & Shin, 2009; Yaman & Çağıltay, 2010).

In her comprehensive review, Paek (2005, p. 17) concludes that overall CBT and PBT "versions of traditional multiple-choice tests are comparable across grades and academic content." Higgings et al (2005) conducted a survey with 219 4th grade students in an attempt to define any probable score differences in reading comprehension between groups, resulting from the test-mode effect; their research revealed no statistically significant differences. Similarly, in the study of Akdemir and Oğuz (2008), 47 prospective teachers in the departments of Primary School Teaching and Turkish Language and Literature took an achievement test, including thirty questions, both on computer and on paper. At the end of the study, it was revealed that there was not statistical difference between the test-takers' scores in line with the test-administration mode. Hence, the researchers mentioned that "computer-based testing could be an alternative to paper-based testing" (p. 123). Hosseini, Abidin, and Baghdarnia (2014) compared reading comprehension test with multiple-choice items administered on computer and on paper; at the end of the study, no significant difference was found. Retnawati (2015) compared the scores of the participants who took paper-based Test of English Proficiency with the ones who took computer-based version of the test as well, and the results revealed that scores in both exam modes were quite similar. Lastly, Logan (2015) aimed to search the students' performance differences up to exam administration mode within the frame of mathematics course. In total, 807 6th grade Singaporean students took the mathematics test with 24 items and the paper folding test either on computer or on paper. The results displayed that there was no significant difference. In contrast, Choi et al. (2003) found out that taking a listening test on computer offered an advantage for the test-takers since they got higher scores compared to a paper-based listening test. Yağcı et al. (2011) at a state university carried out a similar study on this topic. This time participants were 75 vocational school students in the department of business administration. In order to reveal the probable academic success differences among participants, the exam was done in two ways (CBT versus PBT), and at the end, participants' scores were compared. It was found that students who had taken the computer-

assisted exam outperformed. Hensley (2015) carried out a study with 142 students in the department of mathematics at the University of Iowa with an aim to compare the students' test scores taken from paper-based tests and computer-based tests. At the end, it was found that the test scores could not be compared because there was a significant difference between the two test modes. A recent study done by Hakim (2017) with 200 female students whose English language command at B1 level in Saudi Arabia displayed that tests done in two different versions, CBT versus PBT, had statistically significant differences.

Although professional assessment standards attach great importance to the comparability of CBTs and PBTs, there has been little empirical research that examines the impact of technology on the two main aspects of the assessment, which include the concepts of validity and reliability (Al-Amri, 2008; Chapelle, 1998; 1999; 2001; Chapelle & Douglas, 2006). For example, in a recent study, Chua (2012) compared the reliabilities of CBTs and PBTs by using computer- and paper-based versions of the multiple-choice Yanpiaw Creative-Critical Styles test (YBRAINS) and the Testing Motivation Questionnaire (TMQ) with a five-point Likert scale. The findings revealed that the reliability values were close to each other in CBTs and PBTs. However, Chua (2012) stated that the results might have been different if achievement tests had been used in the study since the test takers' motivation, desire to achieve high scores and context of the test might affect the scores. Dermo (2009) also carried out a study with 130 undergraduate students who took online tests. The research had six perspectives such as affective factors, validity, practical issues, reliability, security and learning and teaching. According to the results, it was concluded that taking online tests was regarded as a practical and secure domain by the participants. As for the validity and reliability of online tests, both factors seemed to be appropriate and related to the curriculum. Al-Amri (2008) administered three tests to each participant who took the same test once on computer and once on paper. In order to determine the effect of the testing mode on reliability, he examined the internal consistency (Cronbach's alpha) of CBTs and PBTs and the results indicated that the internal reliability coefficients ranged between .57 and .70, not as high as expected. In order to check concurrent validity of the tests, on the other hand, a correlational analysis was conducted and the results indicated that each PBT significantly correlated with its computerized version. Overall, there was not any significant effect of the test administration mode on the overall reliability and validity of the tests. In another study (Boo, 1997, cited in Al-Amiri), the test administration mode did not have any impact on the reliability of tests. Utilizing an EFL test battery entitled the Test of English Proficiency developed by Seoul National University (TEPS), Choi *et al.* (2003) investigated the comparability between PBT and CBT based on content and construct validation. Although they did not focus on the measurement of course learning outcomes in higher education, their findings supported comparability between the CBT and PBT versions of the TEPS subtests (listening comprehension, grammar, vocabulary, and reading comprehension) in question.

On the other hand, Semerci and Bektaş (2005) conducted a survey about how to improve the validity of web-based tests. In this regard, they collected data from four different state universities (Anadolu, Sakarya, Fırat Universities and METU) in Turkey, where web-based tests were being administered. The researchers sent emails to a total of 45 people at those universities as to collect data for the study, and only 33 of them wrote back. After the data were analyzed, some ways to improve the validity of web-based tests were defined: Digital identities like fingerprint and voice control should be used; teachers should encourage learners to make projects and research; mini-quizzes and video-conferencing can foster learning, so teachers should make use of them in their courses. Within a similar vein, Delen (2015) aimed to focus on how to increase the validity and reliability of computer-assisted assessment. In this sense, optimum item response time for each question was shown on the screen when the participants were busy with answering the exam items, and the findings revealed that

if students were offered optimum item response time, more valid and reliable tests would be achieved than paper-based tests.

Our review of the related literature indicates that although there have been numerous studies that compare CBTs and PBTs in terms of mean scores, there is little research that specifically deals with the criteria of adequate reliability and accuracy of measurement. Wang and Kolen (2001) developed a framework of criteria for evaluating the comparability between CAT and PBT: (1) validity, (2) psychometric/reliability, and (3) statistical assumption/test administration. We assume that these three criteria can also be used to evaluate the comparability between the linear CBTs and PBTs.

1.4. Research questions

To the best of our knowledge, at a time when Turkish institutions of higher education are on the eve of considering the computerized administration of assessments, there is not even a single study that deals with the comparability of computer- and paper-based tests in English language teacher education programs. Thus, the present research grew out of a desire to learn whether the validity and reliability principles of assessment would be influenced by the test administration mode when pre-service English teachers would take an achievement test for their pedagogical content knowledge course. Thus, the following research questions were formulated to guide the present study:

1. To what extent are the results of a paper-based test (PBT) comparable to those of its CBT version?
2. If the PBT in question has satisfied the criteria of adequate reliability and accuracy of measurement, can its CBT version be considered to have equal reliability and accuracy of measurement?

2. Method

The quantitative research model of the study covers the experimental study - a posttest only design. Accordingly, there is no place for pretests in the study, just the posttests are used. After the participants of the study had been randomly assigned to two groups, the control group took the achievement test in a traditional way while the experimental group took the same exam through a computer-assisted system. When the exam was over, both groups were administered a questionnaire adapted to state some background information of participants and their attitudes towards computer-assisted assessment.

2.1. Participants

The participants for this study consisted of a total of 100 student teachers enrolled in *Approaches to ELT* course in the English language teaching (ELT) department at Hacettepe University. They had already been enrolled in three different sections of the course and taking it from the same faculty member before the study started. They were randomly assigned to the experimental and control groups. During the data collection procedure, three participants dropped from the control group because of different reasons. Thus, in the final data analysis, there were 50 (51.5%) student teachers in the experimental group while there were 47 (48.5%) student teachers in the control group. Their ages ranged from 19 to 23 (N = 97, M = 20.64 years, SD = .84). The researchers also collected data about the participants' grade point averages by classifying them into three groups: students who got between 3.50 and 4.00 (N = 3, 3.1%); students who got between 3.00 and 3.49 (N = 54, 55.7%) and students who got between 2.99 and below (N = 40, 41.2%). Furthermore, an independent samples t-test was run to compare participants in both groups in depth in terms of their *computer literacy,* which is based on

the participants' self-perception, *daily use of internet* and *approximate time of starting to use computer*. As seen in Table 1 below, there were no statistically significant differences between groups. More clearly, the prospective ELT teachers in the experimental group (M = 3.44, SD = .675) and in the control group (M = 3.42, SD = .773) showed non-statistically significant difference in their level of computer literacy ($t_{(95)}$= -.098, p = .92). Similarly, the experimental group (M = 3, SD = 1.14) and the control group (M = 3.08, SD = 1.19) did not show statistically significant difference in terms of their daily use of internet and/or computer ($t_{(95)}$ = .359, p = .72). Lastly, the test-takers in the experimental group (M = 1.58, SD = .575) and in the control group (M = 1.66, SD = .668) did not differ significantly according to their approximate time to start to use computer ($t_{(95)}$ = .630, p= .53). Consequently, it can be easily mentioned that participants have had similar features in both groups.

Table 1. Independent Samples T-Test

		N	M	SD	Mean Difference	t	df	p
Computer Literacy	***PBT***	47	3.426	.773	.02	-.098	95	.922
	CBT	50	3.44	.675				
Daily Use of Internet	***PBT***	47	3.085	1.195	.08	.359	95	.721
	CBT	50	3	1.143				
Approximate Time of Starting to Use Internet/ Computer	***PBT***	47	1.66	.668	.03	.630	95	.530
	CBT	50	1.58	.575				

2.2. Data collection procedures and instruments

Although this study was based on the relation between two different exam modes, the data collection procedure included the content-knowledge course and the achievement test (done both on the computer and on the paper). English language teaching departments offer a course titled *Approaches to ELT* in order to get the prospective English language teachers to identify and describe major language teaching methods including *Communicative Language Teaching (CLT)*. During the data collection phase, this course was offered in three different sections, taught by the same faculty member and taken by all students in the department. Prior to the administration of the achievement test, the course instructor devoted a total of nine hours in three weeks to CLT as the first module in the course. During these three weeks, lectures about the development, principles, assumptions and techniques of CLT were given. Furthermore, the classes were supported by a video-demonstration in order to show a typical CLT-based classroom in ELT. Group presentation was also supported since performance-based assessment was of importance in that course. After all instructional activities were completed in the regular teaching sessions, the students were supposed to get the required information about CLT and they needed to take an achievement test as part of the assessment process. In this regard, the participants took a course achievement test either on a paper or on the web in accordance with their group. The test administration mode was the only difference between the groups. That is, all participants were tested with the same questions which included a total of 60-item developed within the scope of sources used in the content-knowledge course by the instructor by focusing on C*ommunicative Language Teaching.*

Before the exam was administered, an item pool had been generated with questions in different formats like multiple-choice, gap-filling and true/false. In order to supply the validity of the test, all items were revised by a professor and a lecturer in the English language teaching department with an

aim to supply face validity of the test. Upon the suggestions given, the necessary measures were taken and 60 items, which included 50 multiple-choice items with four alternatives and 10 gap-filling items, were chosen. Then, the reliability of the achievement test was tested with the piloted study. The piloting group consisted of 9 prospective English teachers who had previously taken the same course *(Approaches to ELT)*.

As the scoring of the achievement test in the present study was dichotomous, split-half reliability method was utilized so as to calculate the internal consistency level of the test. After the piloting group had completed the test, the items were divided in half as *the odd-numbers* and *the even-numbers* to minimize some probable problems that could be caused by fatigue or boredom of test-takers towards the end of the test or by a test becoming gradually difficult (Blerkom, 2009, p. 49; Ravid, 2011, p. 195; Whiston 2009, p. 54). Then, the reliability coefficient was calculated. The results of the test revealed that the achievement test had very good psychometric properties (r = .90) in terms of reliability. As for the split-half coefficient which gives the value belonging to the half of the test, and the Spearman-Brown coefficient, which gives the reliability coefficient of the whole test, the coefficient values are .896 and .902 respectively; these values demonstrate the reliability of the test used in the present study.

Table 2. Split-half Coefficient of the Achievement Test in the Pilot Study

Cronbach's Alpha	Part 1	Value	0.509
		N of Items	30
	Part 2	Value	0.708
		N of Items	30
	Total N of Items		60
Correlation Between Forms			0.821
Spearman-Brown Coefficient	Equal Length		0.902
	Unequal Length		0.902
Guttman Split-Half Coefficient			0.896

Since the items prepared for the achievement test had a high level of correlation coefficient, the test was accepted as reliable, one of the milestones in developing a test. Then, both paper-based and computer-assisted versions of the exam were generated with the same questions that had been piloted before. The computer-assisted version of the test was prepared by using the online platform, ***www.classmarker.com***, which enables teachers, testers and researchers to prepare and administer online tests. Figure 1 shows how items were displayed on the web.

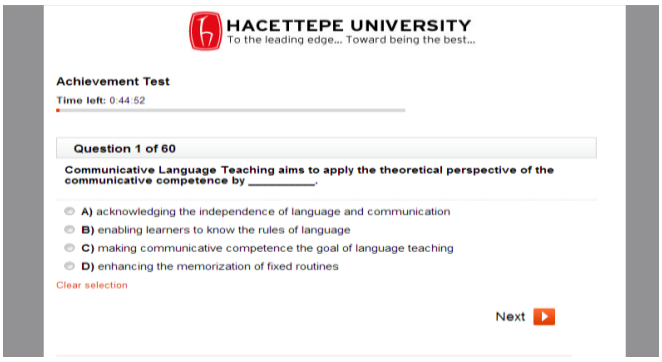

Figure 1. A Sample Item on the System

The test-takers had the chance to move back and forth or skip any question as on paper-based version and to control the allotted time on their screen as seen in Figure 1. As for the gap-filling part, the test-takers were supposed to write the correct or possible answer/s in the given blank. All mandatory or optional answers were mentioned while editing the items on the web (Figure 2). Capitalization or punctuation mistakes were not taken into account for scoring.

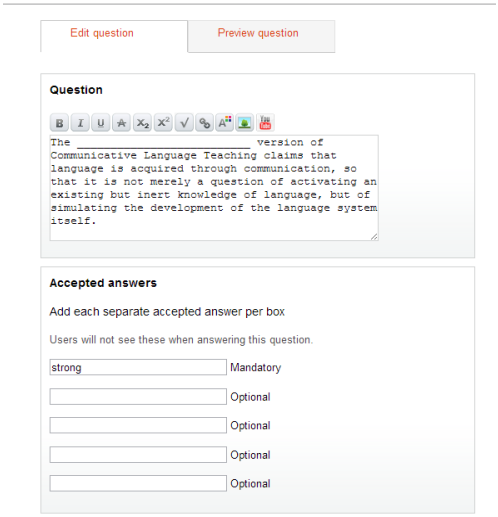

Figure 2. A Sample Item on the System

The participants who took the online version of the test logged into the system by entering their student IDs and passwords that had been previously prepared by the researchers. Once they logged in, they were required to write their full names and email addresses. Before starting the exam, brief information about the exam (the allotted time, the number of questions, the cut-off point for passing the exam, whether they could skip the items or not) was displayed on the screen.

The computer-assisted assessment system offered several benefits both for the test-takers and the teachers: firstly, the test-takers could see their scores on screen just upon completing the exam. In addition, they could get immediate feedback; that is, they saw the correct answer for each item after confirming the question. Secondly, once they achieved 70 or higher points, a certification appeared on screen so as to motivate them. Lastly, the system stored all test-takers' responses, and the researchers reached them whenever needed.

2.3. Data analysis

The present study employed posttest-only experimental research design, a way of gathering quantitative data. All the data were fed into the computer and analyzed by using IBM SPSS 21. At first, the normality level of the data was checked. According to the Kolmogorov-Smirnov test results, it was recognized that the data were not normally distributed ($p < .05$), so further statistical analyses were done in accordance with nonparametric tests. Spearman-brown correlation coefficient (split-half reliability method), the nonparametric equivalent of Pearson product-moment correlation, was calculated so as to reveal the reliabilities of paper-based and computer-based tests. As for the validity values of the tests, Spearman, nonparametric equivalent of Pearson correlation coefficient, was done. Lastly, the probable effect of exam-administration mode over the test-takers' scores was analyzed with Mann-Whitney U test.

3. Results

This part present the results of data analysis based on both descriptive and inferential statistics in order to shed light upon the research questions and aims of the study. A general picture of the participants was given in the Table 3. Accordingly, the number of the test-takers in both groups was almost equal: There were 50 participants (M = 45.86, SD = 7.653) in the experimental group while 47 participants (M = 42.98, SD = 6.479) took the exam in the control group.

Table 3. Descriptive statistics of paper-based and computer-based tests

Tests	*N*	M	Md	Mode	SD	Min	Max
Paper-Based Test	47	42.98	44	38	6.479	32	54
Computer-Based Test	50	45.36	47	49	7.653	29	58

As known, there are two indispensable factors, reliability and validity, for developing and administering a test. In this respect, provided that computers replace papers for exams, the reliability and validity coefficients of both versions should be close and not display any statistical significant difference. Split-half reliability method was utilized for each exam mode, and Table 4 gives the reliability coefficients (r_p =.756, r_c=.903). As a result, it can be mentioned that both exam administration modes have significant level of reliability coefficients. Clearly, there should not be any doubt of reliability once exams are run on computer and/or web-based assessment systems.

Table 4. Reliability coefficients of computer-based and paper-based tests

Test Mode	*r*
Paper-Based Test	.756
Computer-Based Test	.903

Spearman's rank-order correlation, the nonparametric equivalent of Pearson product-moment correlation, was run to assess the relationship between the computer-based test and the paper-based test. While calculating the correlation coefficient in this study, each exam item was assumed a case, not each student. Accordingly, there was a positive correlation between the two variables, r_s = 0.894, n = 60. Table 5 summarizes the results. Overall, there was a strong, positive correlation between the computer-based and paper-based versions of the achievement test. Responses in the paper-based test were correlated with those in the computer-based test. Hence, the concurrent validity of the achievement test was supplied.

Table 5. Spearman's rank-order correlation

			CBT	PBT
Spearman's rho	CBT	Correlation Coefficient	1.000	.894**
		Sig. (2-tailed)	.	.000
		N	60	60
	PBT	Correlation Coefficient	.894**	1.000
		Sig. (2-tailed)	.000	.
		N	60	60

**Correlation is significant at the 0.01 level (2-tailed)

Mann Whitney U test was conducted in order to find out whether there was any statistically significant difference in mean scores between the students who took the computer assisted test and those who took traditional exam. As the dependent variable was continuous and the independent variable was categorical (the subjects were not the same in both groups) and the data related to the exam were not normally distributed, Mann Whitney U Test, non-parametric alternative to t-test, was used (Larson-Hall, 2010, p. 138). The results revealed that there was no significant difference in mean scores of the participants who took the computer assisted test and those who took paper based test, Z_u (954.500), p= .111> .05. More clearly, once students took the exams either on computer or on paper, their performance were not affected in a good or bad way according to test administration mode.

Table 6. Mann-Whitney U test results

Group	N	Mean Rank	Sum of Ranks	Z_U	p*
CBT	50	53.41	2670.50	954.500	.111
PBT	47	44.31	2082.50		

*p>0.05

4. Discussion

The results of the split-half reliability method indicated that both versions of test administration, computer-based and paper-based tests, had high level of reliability coefficients (r_p =.756, r_c=.903). Therefore, it can be deduced that there is not a statistically significant relation between the test-administration modes and the tests' internal consistency levels. Similarly, Chua (2012) found out that computer-based and paper-based tests had close values and revealed the internal consistencies. However, Chua (2012) argues that the results might have been different if achievement tests had been used. As the present study showed, the achievement test did not differ in reliability and internal consistency once it was done both on computer and on paper. In addition, Chua (2012) concludes that computer-assisted tests offer more efficient testing environment for test-takers. This may be because of the fact that computers offer visual cues to test-takers, more authentic exam atmosphere can be achieved via computers, and test-takers have a chance of listening to audio files individually with their earphones. Therefore, computer assisted tests serve in a desired way and test-takers can perform better on computers when they take some kinds of tests such as language skill-based tests (listening items or reading items with some graphs or pictures).

As for the concurrent validity of the tests, Spearman's rank order, the nonparametric alternative to Pearson product-moment correlation, was computed as the data was not normally distributed. According to the data analysis, there was a strong relationship between the computer-based and paper-based tests (r_s = 0.894, n = 60); that is, the aforementioned two exam-administration modes were valid and highly correlated with each other. Similar findings were reported in the studies of Al-Amiri (2008), Choi et a (2003), Dermo (2009) and Siozos et al (2009).

A Mann-Whitney U test was run in order to determine the impact of computers on the scores of the test-takers. The results showed that neither computer-based testing nor paper-based testing affected the success of the test-takers (p=0.111, >0.5). In other words, the paper-based version was found to be comparable to the computer-based one. Though some studies show that CBTs are more challenging than PBTs (Creed, Dennis, & Newstead, 1987; Laborda, 2010) or vice versa (Chin,1990; Dillon, 1994; Yağcı, Ekiz & Gelbal, 2011), some studies supported the findings of the present study (Akdemir &

Oğuz, 2008; APA, 1986; Bugbee, 1996; Choi, et al., 2003; Higgings et al., 2005; Kim & Hyunh, 2007; Logan, 2015; Parshall & Kromrey, 1993; Paek, 2005; Retnawati, 2015; Russel et al., 2003; Stevenson & Gross, 1991; Wang & Shin, 2009; Yaman and Çağıltay, 2010). On the other hand, Choi et al. (2003) mentioned that administering a listening test on computer helped the test-takers get higher scores compared to paper-based listening test since each test-taker could have the chance to listen to the text in a clear way. Similarly, and in contrast to the findings of the present study, Laborda (2010) states in his study that the visual cues presenting on computer create an authentic exam atmosphere for test-takers, and each examinee has an opportunity to listen to a text without being exposed to any external factor that may disturb them. Therefore, students' listening scores can go up. Furthermore, it is really surprising and different from other studies that the students who took computer-assisted tests became more successful in Chin's study (1990), and Yağcı et al. (2011) mentioned that the participants who took computer-assisted tests in their study succeeded 35% higher comparing to the participants who did not. As the related data analysis indicates that there is no exam mode effect, computers can be adopted as alternative testing tools by teachers, because doubts about the affinity of students' scores both on paper and on computer have been eliminated.

Findings of this study revealed many pedagogical implications: this study compared computer-assisted and paper-based modes of the same test. No significant difference between them was found; therefore, computer-assisted exams are said to be alternative forms of traditional tests. In addition, computer-based tests are valid as they serve the aim of the test in a desired way. CBT also gives immediate feedback for incorrect and missing answers (Alderson, 2000; Cohen, 2001; Yunxiang *et al*., 2010), so students have a chance to learn their deficient points and to focus on these areas. Contrary to traditional testing, which takes a long time to announce results and seems a burden for teachers, and most of the time, it is impossible for teachers to give enough feedback for each learner about their mistakes on exam items; computer-assisted testing makes delivering test scores and giving feedback just upon completing the test possible (Alderson, 2000, p. 595), because giving feedback right after any mission done has a crucial and meaningful impact on learning, which is useful for pedagogical purposes (Roever, 2001, p. 85), and assessment is done for both grading students and measuring teaching process, which refers to washback effect. Machine-based or computer-based scoring removes the burden over teachers, and subjectivity on scoring disappears. In addition, special programs are available in order to aid test design, item editing, piloting and having an item pool, which again serves for the principles of effective tests. Moreover, computers offer very rich test content, especially for language tests, and supply the base for communicative language testing (Brown, 2004; Choi *et al.*, 2003; Noyes & Garland, 2008). Clearly, visual cues that are shown on screen during listening tests (Laborda, 2010) or reading tests make the exam atmosphere much more authentic, which is the key component of communicative language testing.

Briefly, technology integrated education, which covers both teaching and assessment procedures, seems to work effectively since there is a growing tendency to utilize cutting-edge technology by the learners. It was succeeded to great extent in teaching, but computers were disregarded as assessment tools. However, recently some leading universities across the world such as Stanford University, Cambridge University and MIT have started to launch computer-based testing since it offers many advantages both for teachers and for learners. Firstly, computer-based, namely web-based, assessment systems give fast and accurate scoring (Alderson, 2000; Cohen, 2001). In other words, computer-based tests reduce human error in scoring (Noyes & Garland, 2008, p. 1369). Secondly, computer-based testing saves time and place; that is, test-takers can reach the test wherever and whenever it is available (Roever, 2001). Thirdly, costs with printing tests in paper are gone down with computers. Lastly, computer-based testing provides authentic materials for testing. For instance, visual cues can be supplied with computers or each test-taker can listen to the text in a clear way once the computer-

based listening test is provided (Choi *et al.*, 2003; Laborda, 2010). Overall, computer technology has many benefits as the related studies display. In addition, the related literature indicates that computers can be used as an alternative for traditional assessment since validity and reliability coefficients are close to each other in both versions, and the test-takers succeed similarly in both modes of assessment.

5. Conclusion

The present study was designed to investigate the impact of computer technology on the two important tenets of assessment, validity and reliability. In this regard, a total of 97 prospective English teachers enrolled in the *Approaches to ELT* course in an ELT department at a major state university in Turkey were chosen as the study group because it was detected that there was no related study with English language teaching programs in Turkey. In line with the purposes of the study, the students were randomly assigned to two groups: the experimental group took the achievement test on computer whilst the control group took the traditional way of assessment. The results indicated that both computer-based and paper-based versions of tests had high level of reliability coefficients, internal consistency and strong relation between each other. Furthermore, it was found that neither computer-based testing nor paper-based testing affected the success of the test-takers, so it can be deduced that the paper-based version can be comparable to the computer-based one.

Without doubt, findings of the present study indicated many pedagogical implications. As it is known, a test has to be based on some principles such as validity, reliability, practicality and washback effect (or backwash effect) and once the test has these principles, it proves the efficiency of it. No significant difference between them was recognized. Moreover, it is known that computers offer very rich test content, especially for language tests, and give immediate feedback to test-takers about their incorrect and missing answers, so the test-takers have a chance to learn their deficient points and to focus on these areas, which is very useful for pedagogical purposes known as washback effect. Machine-based or computer-based scoring removes the burden over teachers, and subjectivity on scoring disappears. In addition, special programs are available in order to aid test design, item editing, piloting and having an item pool, which again serves for the principles of effective tests. Overall, computers can be used as alternatives to traditional testing methods without worrying the core concepts of assessment; instead, the advantages the technology brings to the education should be taken into consideration.

Although the research reached its aims, there were certainly some limitations. First of all, only 97 student teachers in the department of English language teaching were included. More students in the same department from different universities or more students from a variety of departments could have participated in the study, so the results could be easily generalized to higher education system in Turkey. Secondly, the study was only focused on tertiary level students, but high school, secondary school and/ or even primary school students could be covered in these studies. Thirdly, the participants took the computer assisted test only one time, for their midterm exam, but using this system during one semester or during whole year may give more sensitive results. Also, only 9 prospective English language teachers who had previously taken the same course constituted the piloting group, but the number could have been higher. And lastly, computer adaptive testing is the paradigm of the 21st century; however, the researcher couldn't use it, but computer-assisted testing system was utilized.

Now that there are some limitations in the study, further studies can focus on these points. First of all, the researcher used computer-assisted assessment system, but computer-adaptive tests are the contemporary exam modes in this century. Hence, future researchers can make use of these tests. In

addition, students from different departments or from different universities can be included, because as the number of samples goes up, the findings can be generalized in a salient way.

References

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

Alderson, J.C. (2000). Technology in testing: The present and the future. *System, 28,* 53-603. doi.org/10.1016/S0346-251X(00)00040-3

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 and Linguistics, 10*, 22–44.

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

Bennett, R. E. (2003). *Online assessment and the comparability of score meaning*. Princeton, NJ: Educational Testing Service.

Blerkom, M. L. V. (2009). *Measurement and statistics for teachers*. New York, NY: Routledge.

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

Brown, H.D. (2004). *Language assessment: Principles and classroom practices*. White Plains, NY: Pearson Education.

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices.*White Plains, NY: Pearson Education.

Brusilovsky, P., & Miller, P. (1999). *Web-based testing for distance education*. Webnet 99 World conference on the WWW, Hawaii, USA, 24-30 October 1999.

Bugbee, A. C. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education, 28* (3), 282-299.

Chapelle, C. (1998) Construct definition and validity inquiry in SLA research. In L. F. Bachman and A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing* research (pp. 32-70). New York, NY: Cambridge University Press.

Chapelle, C. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19*, 254-72. https://doi.org/10.1017/S0267190599190135

Chapelle, C. (2001) *Computer applications in second language acquisition: Foundations for teaching, testing, and research.* Cambridge, England: Cambridge University Press.

Chapelle, C., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge, England: Cambridge University Press.

Chin, C. H. L. (1990). *The effect of computer-based tests on the achievement, anxiety and attitudes of* grade 10 science students. (Unpublised master's thesis). The University of British Columbia, Vancouver.

Choi, I. C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing, 20*(3), 295-320. doi: 0.1191/0265532203lt258oa

Choi, S. W., & Tinkler, T. (2002). *Evaluating comparability of paper and computer based assessment in a K-12 setting*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Chua, Y. P. (2012). Effects of computer-based testing on test performance and testing motivation. *Computers in Human Behavor, 28*(5), 1580-1586. doi: 10.1016/j.chb.2012.03.020

Cisar, S. M., Radosav, D., Markoski, B., Pinter, R., & Cisar, P. (2010). *New Possibilities for Assessment through the Use of Computer Based Testing*. 8th International Symposium on Intelligent Systems and Informatics, Serbia, 10-11 September 2010.

Cohen, A. D. (2001). Second language assessment. In M. Celce-Murcia (Ed.). *Teaching English as a second or foreign language* (3rd ed., pp. 515-534). Boston, MA: Heinle & Heinle.

Creed, A., Dennis, I., & Newstead, S. (1987). Proof-reading on VDUs. *Behaviour and Information Technology, 6*(1), 3-13. https://doi.org/10.1080/01449298708901814

Delen, E. (2015). Enhancing a computer-based testing environment with optimum item response time. *Eurasia Journal of Mathematics, Science and Technology Education, 11*(6), 1457-1472. https://doi.org/10.12973/eurasia.2015.1404a

Dermo, J. (2009). E-assessment and the student learning experience: A survey of student perceptions of e-assessment. *British Journal of Educational Technology, 40* (2), 203-214. https://doi.org/10.1111/j.1467-8535.2008.00915.x

Dillon, A. (1994). *Designing usable electronic text: Ergonomic aspects of human information usage*. London: Taylor & Francis.

Dunkel, P. (Ed.) (1991). *Computer-assisted language learning and testing: Research issues and practice*. New York, NY: Newbury House.

Flaugher, R. (2000). Item banks. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer, 37-59*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Folk, V. G., & Smith, R. L. (2002). Models for delivery of CBTS. . In C. N. Mills, Potenza, M. T., Fremer, J. J., Ward, W. C. (Eds.), *Computer-based testing: Building the foundation for future assessments, 41-66*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Fulcher, G. and Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. New York, NY: Routledge.

Guzman, E., & Conejo, R. (2005). Self-assessment in a feasible, adaptive web-based testing system. *IEEE Transactions on Education, 48* (4), 688-695. doi: 10.1109/TE.2005.854571

Hakim, B. M. (2017). Comparative study on validity of paper-based test and computer-based test in the context of educational and psychological assessment among Arab students. *International Journal of English Linguistics, 8*(2), 85-91. http://doi.org/10.5539/ijel.v8n2p85

Hensley, K.K. (2015). *Examining the effects of paper-based and computer-based modes of assessment of mathematics curriculum-based measurement*. Unpublished PhD thesis, University of Iowa, Iowa.

Higgings, J., Russell, M., & Hoffmann, T. (2005). Examining the effect of computer-based passage presentation on reading test performance. *Journal of Technology, Learning and Assessment, 3 (4), 3-35.*

Hosseini, M., Abidin, M.J.Z., & Baghdarnia, M. (2014). Comparability of test results of computer based tests (CBT) and paper and pencil tests (PPT) among English language learners in Iran. *Social and Behavioral Sciemces, 98,* 659-667. doi: 10.1016/j.sbspro.2014.03.465

Hughes, A. (2003). *Testing for language teachers*. (2nd ed.). Cambridge, England: Cambridge University Press.

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

Kearsley, G. (1996). The World Wide Web: Global access to education. *Educational Technology Review, 5,* 26-30.

Kim, D. H., & Huynh, H. (2007). Comparability of computer and paper-and-pencil versions of algebra and biology assessments. *Journal of Technology, Learning, and Assessment, 6*(4), 4-30. Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/download/ 1634/1478.

Laborda, J. G. (2010). Contextual clues in semi-direct interviews for computer assisted language testing. *Procedia Social and Behavioral Sciences, 2,* 3591-3595. doi:10.4304/jltr.5.5.971-975

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS.* Abingdon, Oxon: Routledge.

Lilley, M., Barker, R., & Britton, C. (2004). The development and evaluation of a software prototype for computer-adaptive testing. *Computers and Education, 43,* 109-123.

Linden, W. J. (2002). On complexity in CBT. . In C. N. Mills, Potenza, M. T., Fremer, J. J., Ward, and W. C. (Eds.), *Computer-based testing: Building the foundation for future assessments, 89-102.* Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Linden, W. J., & Glas, G. A. W. (2002). *Computer-adaptive testing: Theory and Practice.* NewYork: Kluwer Academic Publishers.

Logan, T. (2015). The influence of test mode and visuospatial ability on mathematics assessment performance. *Mathematics Education Research Journal, 27,* 423-441. doi: 10.1007/s13394-015-0143-1

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design.* Mahwah, NJ: Lawrence Erlbaum Associates.

Madsen, H. S. (1991). Computer-adaptive testing of listening and reading comprehension. In P. Dunkel(Ed.) *Computer-assisted language learning and testing, 237-257.* New York, NY: Newbury House.

McGough, J., Mortensen, J., Johnson, J., & Fadali, S. (2001). *A web based testing system with dynamic question generation.* 31st ASEE/ IEEE frontiers in education conference, Reno, 10-13 October 2001.

Muter, P., Latremouille, S. A., Treurniet, W. C., & Beam, P. (1982). Extended reading of continuous text on television screens. *Human Factors, 24,* 502-508. https://doi.org/10.1177/001872088202400501

Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics, 51*(9), 1352-1375. doi: 10.1080/00140130802170387

Paek, P. (2005). *Recent trends in comparability studies* (Pearson Educational Measurement Research Report 05-05). Retrieved from http://www.pearsonassessments.com/NR/rdonlyres/5FC04F5A-E79D-45FE-8484-07AACAE2DA75/0/TrendsCompStudies_rr0505.pdf.

Parshall, C. G., & Kromrey, J. D. (1993). *Computer-based versus paper-and-pencil testing: An analysis of examinee characteristics associated with mode effect.* Annual meeting of the American educational research association, Atlanta, GA, April 1993.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer based testing.* Verlag, NewYork: Springer.

Ravid, R. (2011). *Practical statistics for educators* (4th ed.) Plymouth, UK: Rowman & Littlefiel.

Retnawati, H. (2015). The comparision of accuracy scores on the paper and pencil testing versus computer-based testig. *TOJET, 14*(4), 135-142.

Roever, C. (2001). Web-based language testing. *Language Learning and Technology, 5(*5), 84-94.

Russell, M., Goldberg, A., & O'conner, K. (2003). Computer-based testing and validity: A look back into the future. Assessment in Education: Principles, Policy & Practice, 10 (3), 279-293. https://doi.org/10.1080/0969594032000148145

Scheerens, J., Glas C., & Thomas, S. M. (2005). *Educational evaluation, assessment, and monitoring: A systemic approach.* Lisse: Swets & Zeitlinger B.V.

Semerci, Ç., & Bektaş, C. (2005). İnternet temelli ölçmelerin geçerliliğini sağlamada yeni yaklaşımlar. *TOJET, 4*(1), 130-134.

Siozos, P., Palaigeorgiou, G., Triantafyllakos, G., & Despotakis, T. (2009). Computer-based testing using "digital ink": Participatory design of a tablet PC based assessment application for secondary education. *Computers & Education, 52,* 811-819.

Stevenson, J., & Gross, S. (1991). Use of a computerized adaptive testing model for ESOL/ bilingual entry/ exit decision making. In P. Dunkel (Ed.) *Computer-assisted language learning and* testing (pp. 223-235). New York, NY: Newbury House.

Stobart, G. (2012). Validity in formative assessment. In J. Gardner, (Ed.). *Assessment and* learning (pp. 233-242). London: Sage.

Texas Education Agency. (2008). *A review of literature on the comparability of scores obtained from examinees on computer-based and paper-based tests.* Retrieved from https://goo.gl/AAdc5o

Tsai, T. H., & Shin, C. D. (2012). A score comparability study for the NBDHE: Paper-pencil versus computer versions. *Evaluation & the Health Professions, 36(*2), 228-239. https://doi.org/10.1177/0163278712445203

Tung, P. (1986). Computerized adaptive testing: Implications for language test developers. In C. W. Stansfield (Ed.). *Technology and language testing* (pp. 9-11). Washington, DC: TESOL.

Wainer, H., & Eignor, D. (2000). Caveats, pitfalls and unexpected consequences of implementing large-scale computerized testing. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer, 271-298*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Wang, H. (2010). *Comparability of computerized adaptive and paper-pencil tests*. [Online: *http://images.pearsonassessments.com/images/tmrs/tmrs_rg/Bulletin_13.pdf*, retrieved in August, 2013].

Wang, H., & Shin, C. D. (2009). Computer-based & paper-pencil test comparability studies. *Test, Measurement and Research Service Bulletin, 9*, 1-6. Retrieved from http://www.pearsonassessments.com/NR/rdonlyres/93727FC9-96D3-4EA5-B807-5153EF17C431/0/Bulletin_9.pdf

Wang, H., & Shin, C. D. (2010). Comparability of computerized adaptive and paper-pencil tests. *Test, Measurement and Research Service Bulletin, 13*, 1-7. Retrieved from http://www.pearsonassessments.com/NR/rdonlyres/057A4A04-9DCB-4B68-9CB0-3F32DDF396F6/0/Bulletin_13.pdf.

Wang, S., 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, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement, 38*(1), 19-49. http://dx.doi.org/10.1111/j.1745-3984.2001.tb01115.x

Ward, W. C. (2002). Test models. In C. N. Mills, Potenza, M. T., Fremer, J. J., Ward, W. C. (Eds.), *Computer-based testing: Building the foundation for future assessments, 37-40*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Whiston, S. C. (2009). *Principles and applications of assessment in counseling (3rd ed.)*. CA: Brooks/ Cole.

Yagcı M., Ekiz. H., ve Gelbal, S. (2011). *Çevrimiçi sınav ortamlarının öğrencilerin akademik başarılarına etkisi*.5th international computer and instructional technologies symposium, Elazığ, Turkey, 22-24 September 2011.

Yaman, S. O., & Cagıltay, N. E. (2010). Paper-based versus computer-based testing in engineering education. *IEEE Educon Education Engineering: The Future of Global Learning Engineering Education,* 1631-1637. doi: 10.1109/EDUCON.2010.5492397

Yunxiang, L., Ruixue, G., Lili, R., Wangjie, Quinshui, Q., & Hefei (2010). *Advantages and disadvantages of computer-based testing: A case study of service learning*. [Online: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5691870, retrieved in July, 2013]. doi: 10.1109/ICISE.2010.5691870

Bilgisayar destekli testler ve klasik testler: Test uygulama yöntemi başarı sınavlarının güvenirliğini ve geçerliliğini etkiler mi?

Öz

Bu çalışma sınavların uygulanma yönteminin, ölçme-değerlendirmenin temel unsurları olan geçerlilik ve güvenirlik üzerinde etkisini araştırmayı amaçlamaktadır. Bu bağlamda Türkiye'de bulunan bir devlet üniversitesinde İngiliz dili eğitimi bölümünde okumakta olan 97 öğretmen adayı çalışmaya katılmıştır. Bütün katılımcılar *İngiliz Dili Eğitimi Yöntemleri* dersine kayıtlı olan öğrencilerdir. Katılımcılar deney grubu ve kontrol grubu olarak ikiye ayrılmıştır. Deney grubunda olan katılımcılar sınava bilgisayar ortamında katılırken kontrol grubunda olan katılımcılar klasik yöntemle sınava girmiştir. Çalışma sonunda sınavların bilgisayar ortamında ya da klasik yöntemle verilmesinin, güvenirlik ve geçerlilik üzerinde etkisi olmadığı göstermiştir. Ayrıca, öğrencilerin bilgisayar ortamında sınava katılmaları başarı puanlarını etkilememiştir. Sonuç olarak, bilgisayar desteği sadece eğitim süresince değil ölçme-değerlendirme aşamasında da kullanılabilir.

Anahtar sözcükler: bilgisayar destekli sınav; klasik sınav; güvenirlik; geçerlilik; İngilizce öğretmen eğitimi

AUTHORS' BIODATA

Hüseyin Öz is an associate professor of applied linguistics and English language teaching at Hacettepe University. He received his MA degree from Middle East Technical University and his PhD degree in Linguistics from Hacettepe University, where he teaches undergraduate and graduate courses in language teaching methods and approaches, research methods, applied linguistics, second language acquisition research, language assessment, and technology enhanced language learning (TELL). He has published widely in various refereed international journals and presented papers in national and international conferences. He has also served on the editorial boards of several national and international publications and is currently the associate and managing editor of Eurasian Journal of Applied Linguistics.

Tuba Özturan is a PhD candidate at Hacettepe University and lecturer in School of Foreign Languages at Erzincan University. She received her BA degree in English Language Teaching from Gazi University and her MA degree in English Language Teaching from Hacettepe University.