Eurasia Journal of Mathematics, Science & Technology Education, 2015, **11**(6), 1457-1472

Enhancing a Computer-Based Testing Environment with Optimum Item Response Time

Erhan Delen
Giresun University, TURKEY

•*Received 9 March 2015*•*Revised 14 May 2015* •*Accepted 25 May 2015*

As technology has become more advanced and accessible in instructional settings, there has been an upward trend in computer-based testing in the last decades. The present experimental study examines students' behaviors during computer-based testing in two different conditions and explores how these conditions affect the test results. Results indicate that some of the psychometric features of a test (reliability and validity) could be enriched on computer-based testing if students are provided optimum item response time. In addition, it was found that providing optimum response time for each item influenced the students in the experimental group to not engage in rapid guessing behaviors. Thereby, students spent a reasonable amount of time answering the questions, which resulted in more reliable and valid scores, aforementioned. Lastly, there was no statistically significant difference in two groups in terms of student performance.

Keywords: Computer-based testing, psychometric features, reliability, validity

INTRODUCTION

Assessment is one of the indispensable parts of the educational process. There are many measurable components to be assessed in education including knowledge, skills, attitudes, and perceptions. Hence, researchers have used numerous methods and techniques to acquire valid (Abedi, 2014; Chou, Moslehpour, & Huyen, 2014; Schatz & Putz, 2006), reliable (Chua, 2012), and meaningful assessment results.

As education has become more advanced in the last decades in various ways, educators and researchers have proposed new approaches for assessment practices in education. For instance, due to computer use in educational settings and a significant interest in distance education, there has been an upward trend in computer-based learning. This trend has also changed the mode of assessment from paper-based to computer-based (Chua & Don, 2013; Hosseini, Abidin, & Baghdarnia, 2014; Weinerth, Koenig, Brunner, & Martin, 2014). This change was necessary because computers and related technologies (e.g., mobile devices) have many affordances for the instruction and assessment process.

Now, we use a comprehensive term to indicate this use: e-assessment. Researchers have used "E" with other terms such as mail (email), book (ebook), and learning (elearning). Now, it is time to Enrich assessments with Electronic formats.

Correspondence: Erhan Delen,
Department of Computer Education and Instructional Technology, Faculty of Education, Giresun University, Güre, Giresun, Turkey.
E-mail: erhan.delen@giresun.edu.tr
doi: 10.12973/eurasia.2015.1404a

ISSN: 1305-8223

Although mobile devices, such as phones and tablets, are also used for e-assessment, most of the researchers have used the terms ‘e-assessment’ and ‘computer-based testing’ synonymously due to dominance of computers in e-assessment (García-Peñalvo, 2008; JISC, 2007).

There are many advantages of using computers when measuring test takers’ performances including accurate scoring (JISC, 2007), dynamic results reporting (Debuse & Lawley, 2014), and tracking students’ behaviors (Brown & Abeywickrama, 2010; Olea, Abad, Ponsoda, Barrada, & Aguado, 2011). In other words, with computerized testing, it is easy to avoid answering and scoring errors. For example, some students make mistakes when marking answers on the bubble sheets. In computer-based testing, students are only required to select the correct answer for a single question on the screen, which is very straightforward and error free. Additionally, in true-score theory (Allen & Yen, 1979), it is important to determine test takers’ actions during a test and influences of these actions on the test results. When students are tested on computers, it is possible to monitor, track, and log their answering behaviors with programing techniques. This logged data can be analyzed and reported along with the students’ scores. Essentially, if we know more about the testing process and how test takers behave during the test (not only their scores), we may explain the results more accurately in different ways. For example, we can explore the impacts of rapid guessing behaviors on test scores by monitoring the testing process.

State of the literature

- There has been an upward trend in computer-based testing in the last decades. Many known tests have changed their format from paper-pencil to computer-based in order to benefit from technology for assessment purpose.
- There are many advantages of using computers when measuring test takers’ performances including accurate scoring, dynamic results reporting, and tracking students’ behaviors.
- Item response time and rapid guessing behaviors have been studied by researchers in order to interpret the testing results more accurately.

Contribution of this paper to the literature

- In this experimental study two different computer-based testing environments were designed and tested.
- The benefit of scaffolding students by providing them optimum item response time was the main focus of this study.
- It was showed that the psychometric features of a test could be enriched when students are provided additional features on a computer-based testing environment.

The potential benefits and barriers of computer-based assessments should be explored with empirical studies to provide reasonable confidence for tests on computers (Jeong, 2014; Schatz & Browndyke, 2002). These studies would provide valuable information about the behaviors of test takers during testing. In the current study, a new e-assessment approach is used and tested by tracking and examining students’ behaviors during computer-based testing. The main goal was to observe students’ behaviors during testing in two different conditions and explore how conditions affected the test results.

LITERATURE REVIEW

There are many factors in an assessment process that affect the quality of the test scores. Test makers need to take those factors into consideration when interpreting the scores and conducting upcoming tests. Test takers’ behaviors during a test may offer valuable information about the item answering process for each question and the overall test. If we know how students approach each item, we may be able to improve the quality of questions in various ways, such as reliability, validity, and discrimination (Schatz & Browndyke, 2002). In a computer-based testing environment, we can easily obtain valuable information about the testing process, which could not be obtained from a paper-pencil type test.

Computer-based testing

Computer-based testing has become prevalent and it offers numerous advantages (Clariana & Wallace, 2002; Schatz & Putz, 2006). For instance, "computer-based tests can provide new objective, valid and reliable measures for traditional or new competencies" (Wirth, 2008, p. 246). Additionally, computers enable testing to occur anytime and anywhere (Jeong, 2014), which increases test takers' motivations (Chua & Don, 2013). Computer-based testing also provides financial advantages by requiring fewer human resources and less paperwork (Kaya & Delen, 2014; Schatz & Browndyke, 2002).

Standardized tests, such as Test of English as a Foreign Language (TOEFL) and Graduate Record Examinations (GRE), are the tests that changed their assessment format from paper-based to computer-based. Additionally, the Programme for International Student Assessment (PISA), which is a study that is conducted in many countries, will be administered via computer in 2015 (Weinerth et al., 2014).

Computer-based testing does not only mean moving questions from paper to screen (Jawaid, Moosa, Jaleel, & Ashraf, 2014; Lee, 2009). Essentially, computer-based tests should use advantages of technology as much as possible and support the quality of the tests in various ways (Wirth, 2008). We can use computers for assessment purposes and facilitate the work of both test makers and students with numerous advantages (Kaya & Delen, 2014; Schatz & Browndyke, 2002). However, these advantages may be guaranteed as long as test takers do not have a negative experience with computer-based testing.

Because today's students are digital natives (Prensky, 2001), and they are exposed to technology in many ways, we can benefit from this advantage in instructional settings (Clariana & Wallace, 2002). Schatz and Browndyke (2002) pointed that "computer and Internet technologies have moved from 'emergent' status to 'current' acceptance" (p. 405). Hence, it would not become an issue for test takers to be tested on computers as long as the medium of questions' presented is appropriate (Deboer et al., 2014; Weinerth et al., 2014). Similarly, Ricketts and Wilks (2002) stated that "computer-based assessment is generally acceptable to students" (p. 478). It is important to note that although students do not resist to be assessed on computers, they may still need motivation and encouragement during the testing process.

There is a substantial body of research that found superiority for computer-based tests in various aspects when compared to pencil-paper tests (Charman & Elmes, 1998; Clariana & Wallace, 2002; Sly & Rennie, 1999). For example, Chua and Don (2013) found in their study that students had significantly more testing motivation scores on a biology test in a computer-based test than the paper-based test. Researchers have also studied test takers' perceptions about computer-based testing. In a study by Hosseini et al. (2014), it was found that students had more positive attitudes towards computer-based tests when compared to paper-based tests. Another study among postgraduate students reported that 61.8% preferred computer-based tests compared to paper-based tests (Jawaid et al., 2014). In terms of test scoring, although some studies reported equivalent results for both computer and paper-based tests (Mason, Patry & Berstein, 2001), there are some other studies that claimed the computer-based test as a reason for better scoring (DeAngelis, 2000). For example, Clariana and Wallace (2002) conducted a study with undergraduate students to assess their learning in the Computer Fundamentals course with both computer and paper-based tests. The study results showed that students performed better in the computer-based test compared to the paper-based test.

Psychometric features

One of the most important goals for test makers is to ensure the psychometric features of a test. The results need to be reliable and valid. In addition, item discriminations and item difficulties need to be in a reasonable level. There are many factors that may affect the psychometric features of a test. Students' approach to the test is one of those factors.

Especially in low-stakes tests, students disregard the test result. These kinds of tests do not have substantial consequences for test takers and thereby cause lower test motivation and test scores for students (Kong, Wise, & Bhola, 2007). Similarly, Chua and Don (2013) stated that "achievement test may be influenced by context of test, for example, motivation and willingness of the participants to achieve higher scores in the tests" (p. 1894). Hence, test makers could find ways to increase students' commitment with well-designed testing environments and observe them during testing to ensure the quality of test scores in different perspectives (Lee & Jia, 2014). However, paper-pencil tests are limited in their ability to provide useful data (Kong et al., 2007). That's why computers could be used for assessment purposes. Tests' reliability could be increased by changing test format to computer-based because computers"are better able to provide precise control over the presentation of test stimuli" (Schatz & Browndyke, 2002, p. 397). In addition, Russell, Goldberg, and O'connor (2003) note that computer-based tests provide valid results when they are well designed. Computer use in testing provides beneficial data about student behaviors, such as item response time, which may give researchers clues about the statistical issues of test results (Kong et al., 2007).

Item Response Time

The item response time during testing is an important aspect to understand students' attitudes and items' quality. However, test makers do not usually pay attention to response time (Schatz & Browndyke, 2002). There are several factors that may affect response time, such as item difficulty and item length. It is important to note that the response time is also influenced by the commitment levels of test takers (Lee & Jia, 2014). Specifically, when a student does not have enough knowledge to answer one question or does not take the test seriously, he/she will exhibit rapid guessing behavior (Kong et al., 2007). It is important to note that "The accuracy of such rapid guesses is typically at or near the chance level, as the responses are essentially random" (Lee & Jia, 2014, p. 2). Students' behaviors, such as rapid guessing, would influence the test results in various ways such as reliability and validity (Wise & Kong, 2005; Wise & DeMars, 2006). There are several studies that propose new approaches for improving test's measurements by focusing on response time (see Kong, Wise, Harmes, & Yang, 2006; Wise & DeMars 2006; Wise, Bhola, & Yang, 2006).

Although, item response time would provide us essential information to understand and explore test results from different aspects (Bulut & Kan, 2012), it is neither easy nor practical to obtain students' response time for each question in paper-pencil type tests (Kong et al., 2007). In addition, Lee and Jia (2014) noted that "the paper provides a way to address the issue of rapid-guessing behavior" (p. 21). On the other hand, we can obtain much more useful information than the response time if we use computer-based testing (Abedi, 2014; Kong et al., 2007; Schatz & Browndyke, 2002; Schatz & Zillmer, 2003; Weinerth et al., 2014; Wirth, 2008). Current technology provides us numerous opportunities in educational assessment (Abedi, 2014; Adesina, Stone, Batmaz, & Jones, 2014; Jeong, 2014), such as tracking students' behaviors during a computer-based test.

Computer-based assessment is perhaps the best way of understanding what students do during an assessment. If researchers knew more about students' approaches to items in testing, they would be aware of students' needs and scaffold them properly to have test scores in better quality.

Instructional scaffolding and self-regulation

Students, in general, are in need of instructional scaffolding to acquire sufficient knowledge and skills (Vygotsky, 1978). Thus, they may need some type of support to demonstrate their test performance accurately when they are assessed on computers. It is important to note that students' personal characteristics have influence on how they benefit from computer-based testing. Clariana and Wallace (2002) name this impact the "test mode effect" (p. 593). Hence, computer-based assessment environments could be enriched with embedded additional features to reinforce students' assessment process (Abedi, 2014). For instance, optimum response time for each item could be showed on screen in order to support students' time management during testing and encourage them not to answer questions rapidly.

Being a self-regulated learner is very important in order to achieve in learning and testing. According to Zimmerman (1989) there are many strategies that a self-regulated learner could use during the learning and evaluation process. Self-evaluation is one of those strategies that require monitoring self-improvement (Pintrich, 1999). Students could be supported to use this strategy by embedding new features to computer-based and online environments (Delen, Liew, & Willson, 2014). Computer-based testing environments could have many features that facilitate self-evaluation for the test takers. For instance, immediate feedback could be given to the learner when the student answers a question.

Each student may have different answering strategies during a test. For instance, a student could underline the important sections when reading a question while another student takes short notes. Hence, computer-based assessment environments need to be also examined regarding to answering strategies. For example, students' mouse movements and item selection tactics could be monitored and interpreted based on the test scores. It is important to note that the use of input devices such as keyboard or a computer mouse may affect the validity of the data collected (Wirth, 2008).

In essence, studies suggest that computer-based testing has many advantages and may be administered in numerous ways to enhance students' individual performance and test results, in general (Charman & Elmes, 1998; Chua & Don, 2013; Clariana & Wallace, 2002; Sly & Rennie, 1999). Hence, the purpose of this experimental study was to examine whether providing optimum item response time to students in the computer-based testing environment enhances the psychometric features of the test. Additionally, relations between total time, mouse movement on items, item change and test performance in computer-based assessments were examined.

Research questions

Two broad research questions were examined in this study on computer-based testing.

RQ 1: Do psychometric features of a test change when providing optimum response time for each test item in an enhanced computer-based testing environment?

RQ2: Do students' response behaviors and performance change in the enhanced computer-based testing environment compared to a common computer-based testing environment?

METHODOLOGY

Research design

To address the two research questions, the present study used a cross-sectional experimental design with one control group using a common computer-based testing environment and one experimental group using an enhanced computer-based testing environment.

Participants

Participants were freshmen students from the Department of Primary Education at a university located in the north eastern region of Turkey. Students were from three sections of Computer-I course. The sections were randomly assigned to the control group (two sections) and the experimental group (one section). Students were visited in a computer lab during their course and asked to participate in the study. The course instructor awarded participants with five bonus points for participating. Additionally, the researcher provided a 1-week free lunch ticket to students in both control and experimental groups who were in first place based on the performance test. Data from one participant were excluded from the analysis due to technical failure during the study. As a result, a total of 94 students participated, 58 students were assigned to the control group and 36 students were assigned to the experimental group.

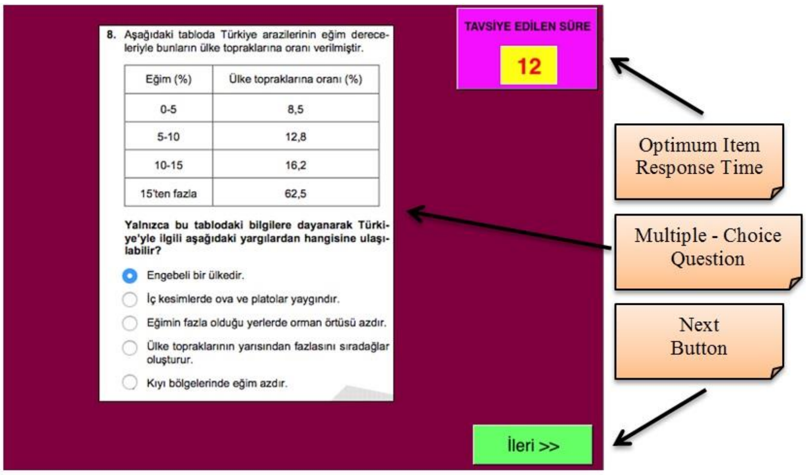
Instruments

In this study, data were collected using three primary measures: a) a geography performance test with 24 multiple-choice questions, b) students' individual geography test scores from the Undergraduate Placement Examination (UPE) in the year of 2014, and c) students' behaviors, which were tracked and logged by the computers during testing. The geography performance test was a part of the UPE from a previous year. In Turkey, all high school graduates need to take the UPE, which is a standardized test, to apply to undergraduate programs in universities. The geography test was selected to examine students' performance because students who apply for the Primary Education Department need to answer geography questions in the UPE. Hence, participants were familiar with the test content. Students' individual geography test scores from the UPE in the year of 2014 were considered as a pretest data to test students' readiness. Students' UPE scores were obtained from the official website, which contains test takers' scores for each subtest. Details about logged data of students' behaviors will be discussed in the next section.

Design and development of common and enhanced computer-based test environments

Two different computer-based testing environments were developed for this study. There were similarities and differences between these testing environments in terms of their features. The first environment was named common computer-based because the environment consisted of questions in electronic formats and students answered the questions on a computer screen similar to paper-based tests. The second environment was named enhanced computer-based because the environment was enriched with a newly added feature (optimum item response time), which was the main difference between the two platforms.

In the enhanced computer-based testing environment, the main distinction was a feature called Optimum Item Response Time. This feature aimed to provide an optimum response time for each question on the screen, when the question appears



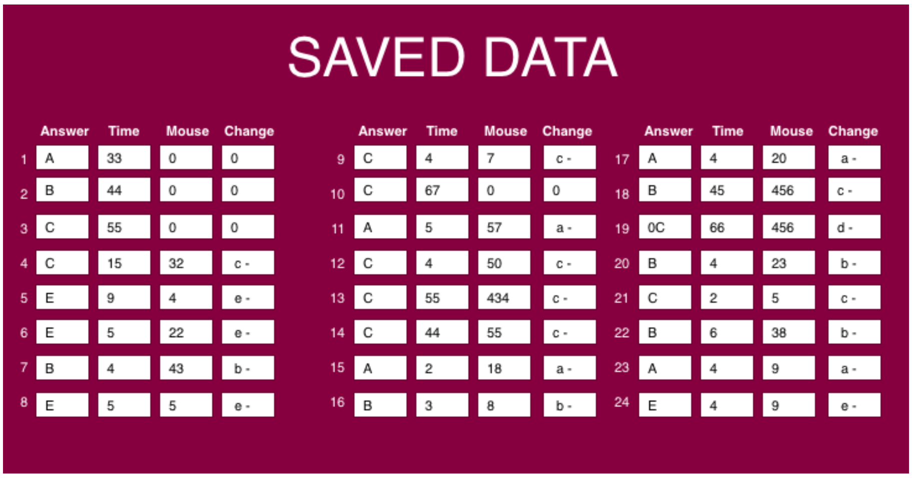
8. Aşağıdaki tabloda Türkiye arazilerinin eğim dereceleriyle bunların ülke topraklarına oranı verilmiştir.

Eğim (%)	Ülke topraklarına oranı (%)
0-5	8,5
5-10	12,8
10-15	16,2
15'ten fazla	62,5

Yalnızca bu tablodaki bilgilere dayanarak Türkiye'yle ilgili aşağıdaki yargılardan hangisine ulaşılabilir?

- ● Engebeli bir ülkedir.
- ○ İç kesimlerde ova ve platolar yaygındır.
- ○ Eğimin fazla olduğu yerlerde orman örtüsü azdır.
- ○ Ülke topraklarının yarısından fazlasını sıradağlar oluşturur.
- ○ Kıyı bölgelerinde eğim azdır.

Figure 1. The enhanced computer-based testing environment



SAVED DATA

	Answer	Time	Mouse	Change		Answer	Time	Mouse	Change		Answer	Time	Mouse	Change
1	A	33	0	0	9	C	4	7	c -	17	A	4	20	a -
2	B	44	0	0	10	C	67	0	0	18	B	45	456	c -
3	C	55	0	0	11	A	5	57	a -	19	0C	66	456	d -
4	C	15	32	c -	12	C	4	50	c -	20	B	4	23	b -
5	E	9	4	e -	13	C	55	434	c -	21	C	2	5	c -
6	E	5	22	e -	14	C	44	55	c -	22	B	6	38	b -
7	B	4	43	b -	15	A	2	18	a -	23	A	4	9	a -
8	E	5	5	e -	16	B	3	8	b -	24	E	4	9	e -

Figure 2. Sample data logged by computer

on screen (see Figure 1). Detailed information about this feature will be given in the procedure section.

The testing procedure was straightforward in both circumstances. Essentially, students were supposed to choose the correct answer for each question, which was given on a separate page on the computer screen. Each question was viewed separately on each computer screen as Ricketts and Wilks (2002) suggested. Once students chose one of the five options in the multiple-choice item, a next button appeared to go to the next question. Students were not allowed to view the previous question once they went on to the next question. Students' behaviors (i.e., item response time, mouse movement, and item selection) were tracked and logged by the computer in both groups. Figure 2 shows a sample data sheet from one student obtained for the study. When students use a new environment, they may face some challenges. Hence, the researcher kept the computer-based testing environments very simple to avoid cognitive challenges as Deboer et al. (2014) suggested.

Procedures

The study was carried out in two successive phases. In the first phase, students in the control group took the geography test on the common computer-based environment. Before the test started, the author instructed students on the testing procedures. There was no time limit for the test. Each student entered the environment with their names and a passcode to start the test. During the test, each question was displayed in a single window on the screen. Once students answered all the questions, the environment was closed automatically. Aforementioned, students' behaviors during the test were tracked and logged by computers.

Once the control group finished the test, internal consistency reliability was calculated and 8 questions were excluded from the test to increase the reliability to a reasonable level. As a result the reliability was Cronbach's α = .614 for the remaining 16 questions. Cronbach's alphafor the test was moderate. However, Hair, Black, Babin, & Anderson (2010) note, "The generally agreed upon lower limit for Cronbach's alpha is .70, although it may decrease to .60 in exploratory research." (p. 125). Given the current study is exploratory in nature, 16 items were used.

Next, students' response times for each question was obtained from the computer logs and analyzed. For each item, an optimum response time was calculated using students' response times from the students who answered the question correctly. An average response time was calculated to find an optimum response time after eliminating outliers with the generalized ESD method (Rosner, 1983).

In the second phase, students in the experimental group took the same test with the enhanced computer-based testing environment. A similar procedure to the control group was followed for the experimental group. However, in this case, an optimum time for each question, which was the average response time from the control group's data, was presented in each question (see Figure 1). The optimum response time was displayed once the question started and a count down began. The optimum response time was used to scaffold the testing process. It is important to note that it was not mandatory to answer the questions in a suggested time. The main reason of using the feature was to help students to answer the questions in a reasonable time. As in the control group, computers also logged students' behaviors in the experimental group.

Data analysis

The data were analyzed by using IBM SPSS 20 statistical software. Means, frequencies, and other descriptive statistics were calculated and reported for test items. To examine whether the test's psychometric features and/or students' behaviors and test performances differed across the two conditions, comparison and correlation analyses were conducted based on the logged data and students' test results. UPE – 2014 test scores were used as a covariate when students' performances were compared with an ANCOVA test.

RESULTS

Descriptive statistics are presented in Table 1 regarding the variables investigated, followed by findings for the two research questions. Total Time refers to duration in minutes that test takers spent to answer the 16 items, and Total Test Score refers to the number of total correct responses on the test. Total Mouse Movement is a numeric value that refers to total mouse movements on items during the test. Total Item Response Clicks indicates the number of times test takers clicked on response options when answering a test question. UPE-2014 Scores are students' geography test scores from the UPE that they took before applying to the university.

Effects of using optimum response time on psychometric features of a test

The first goal of this study was to examine the test's psychometric features in two different computer-based conditions with the RQ 1: Do psychometric features of a test change when providing optimum time for each test items in an enhanced

Table 1. Means and standard deviations of major variables

	Mean	SD	95% CI around the mean	
			Lower	Upper
Control Group [a]				
Total Time	13.54	2.19	12.94	14.13
Total Test Score	9.76	2.82	9.04	10.46
Total Mouse Movement	17600	9123	15201	1999
Total Item Response Clicks	19.64	5.33	18.41	21.12
UPE-2014 Score	11.55	2.74	10.83	12.27
Experimental Group [b]				
Total Time	15.29	3.75	14.10	16.58
Total Test Score	9.72	2.95	8.74	10.65
Total Mouse Movement	20912	12368	16728	25097
Total Item Response Clicks	18.44	4.73	17.09	20.05
UPE-2014 Score	10.23	2.92	9.25	11.22

[a] *n=58 (9 male - 49 female)*
[b] *n=36 (8 male - 28 female)*

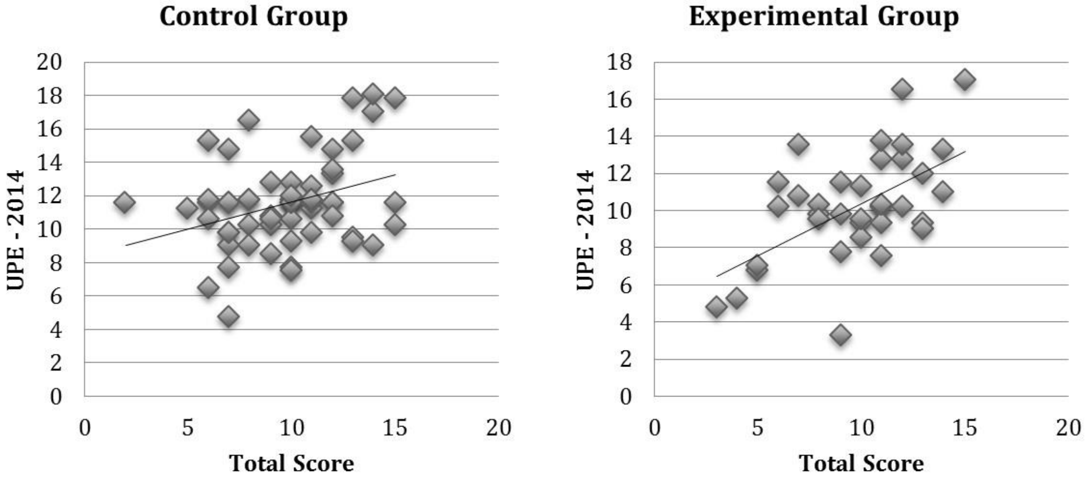


Figure 3. Correlations between UPE – 2014 and computer – based testing

computer-based testing environment? The test's psychometric features was tested and compared in five domains including score reliability, score validity, test difficulty, item discrimination, and item difficulty.

To compare score reliability in two conditions, Cronbach's α method was used. The internal consistency was α = .614 for the control group, whereas the internal consistency was α = .673 for the experimental group. This difference suggests that a computer-based testing tool revealed more reliable test scores for the students when students were provided with the optimum response time for each test item.

In terms of score validity, concurrent validity was taken into consideration to compare two conditions in the study. In concurrent validity, individuals' scores are compared with their scores from a previously taken test. Students' scores in UPE-2014 were correlated with their performance scores on the computer-based test. As Figure 3 shows, the correlation coefficients were $r(56) = .334$, $p < .05$, $r^2 = 11\%$ for the control group, and $r(34) = .566$, $p < .05$, $r^2 = 32\%$ for the experimental group. These results indicate that the enhanced computer-based test environment provides more valid scores, albeit weak, when compared to the control group.

The test difficulty levels were calculated based on the average scores in both groups. According to the students' total scores, the test difficulty levels were quite similar in the control and the experimental group, .62 and .61, respectively.

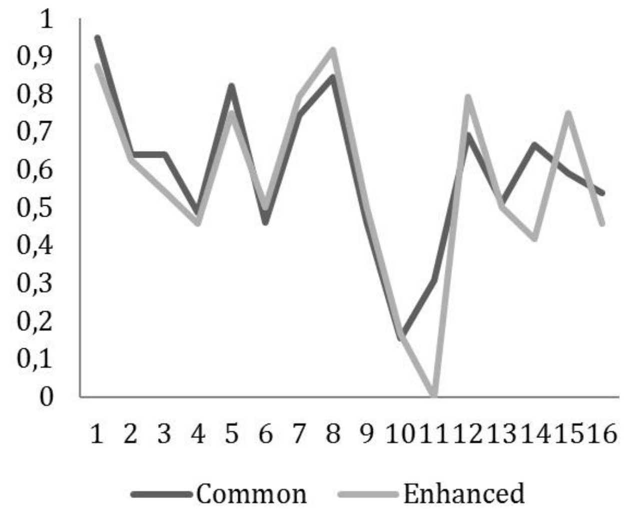


Figure 4. Item Difficulty Indices

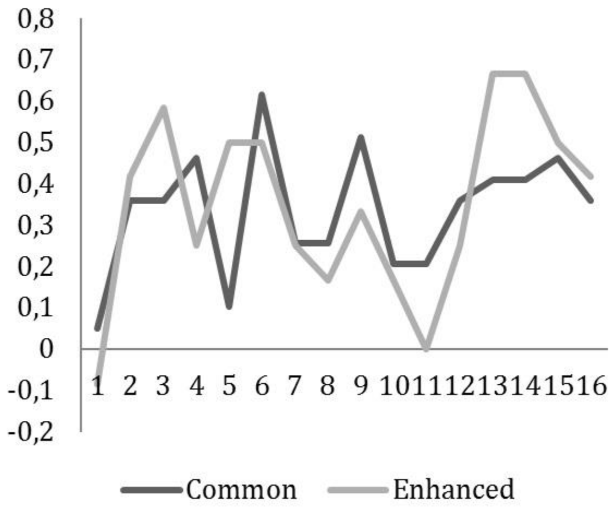


Figure 5. Item Discrimination Indices

Item difficulty and item discrimination indices were calculated for all items based on students' responses in two conditions. Upper and lower 27% of students' scores were considered during the computation. On average, there were no statistically significant differences between the control group and the experimental group in terms of item difficulty (means were .59 and .57, respectively, p>.05) and item discrimination (means were .34 and .35, respectively, p>.05). Figure 4 and Figure 5 show the item difficulty and item discrimination indices.

Effects of using optimum response time on student behaviors and test performance

The second goal of this study was to compare test takers' behaviors during testing and test performances with the RQ2: Do students' response behaviors and performance change in the enhanced computer-based testing environment compared to a common computer-based testing environment? To test whether students acted differently on two conditions, several independent t-tests were conducted to compare Total Time, Total Mouse Movement, Total Item Response Clicks, and Total Test Score.

Time is an essential factor in testing. It is aimed to get more reliable testing results in a reasonable amount of time. In the current study students' time spent was calculated and examined for 16 items. Based on an independent samples t-test, results indicate that students in the experimental group spent more time during testing than students in the control group, $t(50.926) = -2.555$, $p < .05$, $d = .57$. According to this result, it can be concluded that providing optimum response time for each item in the enhanced computer-based testing environment influenced the students in the experimental group to not engage in rapid guessing behaviors.

Mouse movements were tracked in both groups during testing process. Whenever students moved the mouse cursor on the presented item on the screen, the computer converted this movement to a numerical value and saved the value to the database. In terms of total mouse movement, students in the control group did not differ from the students in the experimental group $t(59) = -1.389$, $p > .05$. In addition, correlations were conducted to examine the relationships between students' test scores and their total mouse move on the items. Results indicate that there was no statistically significant linear association between mouse move and student performance in the experimental group $r(34) = .124$, $p > .05$, whereas there was a statistically significant relationship between students' mouse move and their test performance of those in the control group $r(56) = .339$, $p < .05$.

Table 2. Analysis of covariance for test scores by UPE-2014

Source	SS	df	MS	F	p
UPE-2014	138.884	1	138.884	20.419	.000
Group	6.144	1	6.144	.903	.344
Error	618.959	91	6.802		
Total	9684	94			

Students' answer choice selection behaviors were also tracked. That is, the number of different answers they chose before moving on to the next question. For the control and experimental group, the average number of answer choices was approximately 20 and 18, respectively. According to an independent samples t-test, there was no statistically significant difference among the groups, $t(92) = .274$, $p > .05$. This suggests that when students answered the questions, they answered it by selecting the answer approximately only once. In other words, students' confidences in answering the questions were similar.

Lastly, students' test scores were taken into consideration to compare group performances by analyzing number of correct items on the test. After checking and ensuring the homogeneity of regression slope assumption, a one-way analysis of covariance on test scores was conducted, with UPE scores as the covariate to see whether test conditions influenced students' scores. As shown in Table 2, there was no statistically significant difference between students' scores in both groups when controlling for their readiness level using the UPE scores, $F(1,93) = .903$, $p = .344$. According to this finding, it could be concluded that students performed equally in both environments. In other words, providing optimum response time on a computer-based testing environment did not have a statistically different effect on student performance.

DISCUSSION

There are many studies that try to transfer paper-pencil tests to computer-based testing environments in order to benefit from the computerized assessment process. It is important to note that, test takers should not be influenced negatively due to testing environment change when they are asked to take a test on the computer. In other words, outcomes need to be equivalent in terms of student performance. This study aimed to propose a new approach for computer-based testing to enrich the quality of a computerized test in several perspectives compared to a common computer-based testing environment. One goal was to examine how a test's psychometric features could be influenced when test takers are provided with the optimum item response time on the screen. According to results, it was found that when students are scaffolded with the optimum item response time feature, the test revealed more reliable and more valid test scores without causing any differences in students' performances. These findings compliment Schatz and Browndyke (2002), Clariana and Wallace (2002), and Russell et al. (2003), who concluded that computer-based tests would help to increase score reliability and score validity when tests are well designed.

In a similar vein, Wirth (2008) suggests computer-based tests provide numerous opportunities for designing new types of items and tests. Thus, it is important to benefit from computer features and find new ways to measure student outcomes to increase the quality of the testing process and test scores. Moreover students' motivation is an important aspect for score reliability and score validity. As mentioned above, students present less test motivation when it is a low-stakes test (Kong et al., 2007). Hence, it may be concluded that when useful features are embedded into computer-based testing environments, these features can affect students' approach to the test positively (Chua, 2012).

In addition, test difficulty, item discrimination, and item difficulty levels remained the same in two conditions. The levels were in a reasonable range. This result also shows that we can obtain more reliable and valid scores from a test without changing the level of the difficulty and discrimination of test items.

In terms of total time, it was found that students in the experimental group spent more time than students in the control group. In other words, test takers did not engage in rapid guessing behaviors in the experimental condition. Essentially, the newly added feature influenced students to spend a sufficient and reasonable amount of time when answering the questions in the experimental condition. This finding resonates with Wise et al. (2005) who stated that rapid guessing behaviors influence the test results in various ways. According to Kong et al. (2007), students tend to answer test items very quickly when they do not take the test seriously. Hence, from this finding, it may be suggested that by supporting students with different added features in computer-based testing environments, students would take the test more seriously.

Another factor that was observed in this study was students' mouse movement on the screen when answering the questions. In paper-based tests, test takers usually use their pencils to follow up the question when they read. But, in computer-based testing this can be done with the mouse cursor. Hence, students mouse movements were tracked and analyzed based on the cursor. The result suggests that students in both groups used the mouse cursor equally. In addition, there was a statistically significant correlation between mouse movement and test performance in the control group. Whereas, there was no statistically significant correlation in the experimental group. This result may be explained by the total time. In other words, students in the control group answered the questions faster than other students. Thus, students may have benefited from the mouse cursor more sufficiently than the students in the experimental group because they read the questions very quickly and needed the mouse cursor to follow the text in the items. In the experimental group, however, students were not in rush because they were provided optimum response time, and they had chance to self-regulate themselves. Hence, their mouse movement was less important to comprehend the questions.

Students' answer choice selection was also tracked. According to the results, it could be concluded that students selected an answer choice approximately one time once. Meaning that, students did not change their selection before going to the next question. The two groups did not statistically differ from each other in terms of selecting answer choices. This may have happened due to the type of the test. Low-stakes tests do not have substantial consequences for test takers (Kong et al., 2007), and thereby they may not force themselves to review all the answer choices to pick the correct option.

Lastly, students' performances were compared between the two conditions to see whether test environments influenced students' test results. As expected, there was no statistically significant difference in terms of the test results. This finding was important because it is essential to improve a test's quality without affecting students' performances, aforementioned. The aim should be to enrich the test in various perspectives, while keeping students' performances the same. Computerized tests are not supposed to increase test performance; instead, computerized tests need to scaffold students during the test and provide opportunities to reveal more reliable and valid scores.

CONCLUSION

Computer-based testing seems to be growing more popular because there is an emergent interest in online and computer-based learning in universities (Peat & Franklin, 2002). Thus, online testing tools (mostly with computers) have also become widespread (Deboer et al., 2014; Ogletree, Ogletree, & Allen, 2014). This trend continues to increase as unique benefits of computers in assessment emerge. Computers provide many affordances for this new horizon. Hence, it is important to design and test new computer-based testing tools, which provide contemporary approaches to the testing process. There are many advantages of using computers when measuring test takers' performance including accurate scoring (Clariana et al., 2006; Russell et al., 2003; Schatz & Browndyke, 2002), immediate results (Debuse & Lawley, 2014), and tracking students' behaviors (Brown & Abeywickrama, 2010; Olea et al., 2011). Subsequently, computer-based tests should be designed to use advantages of technology and improve the quality of the tests in various ways such as usability, reliability, and validity (Wirth, 2008). This study focused on using an optimum item response time feature in a computer-based test. Based on the findings, it may be suggested that when test takers are scaffolded with optimum response time information, the test scores become more reliable and valid. In essence, effective use of computers can make substantial contributions to educational measurement and evaluation (Dindar, Yurdakul, & Dönmez, 2013). Future studies should explore the potential benefits and barriers of computer-based assessments to provide reasonable confidence for computerized tests (Jeong, 2014; Schatz & Browndyke, 2002).

LIMITATIONS

Although the study findings suggest providing optimum response time to students in computer-based testing to enrich the psychometric features of the test, study results need to be interpreted with considering several limitations. The sample size of the study is small for a psychometric study. In addition, the study was not a part of an official course. Thus, the results might have been influenced because of using a low-stakes test. Reliability levels of the tests seem quite low. Hence, more reliable tests could be applied in similar studies. Moreover, due to eliminating some items from the test, the content coverage could not be guaranteed. Lastly, using a single subject limits the generalizability of this study. Future studies could focus on a high-stakes test with a large group of participants with different subjects to test the benefit of scaffolding students with optimum item response time on computer-based testing.

REFERENCES

Abedi, J. (2014). The use of computer technology in designing appropriate test accommodations for English language learners. *Applied Measurement in Education*, *27*(4), 261-272. doi: 10.1080/08957347.2014.944310

Adesina, A., Stone, R., Batmaz, F., & Jones, I. (2014). Touch Arithmetic: A process-based Computer-Aided Assessment approach for capture of problem solving steps in the context of elementary mathematics. *Computers & Education*, *78*, 333-343. doi:10.1016/j.compedu.2014.06.015

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

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

Bulut, O., & Kan, A. (2012). Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Egitim Arastirmalari-Eurasian Journal of Educational Research, 49*, 61-80.

Charman, D. & Elmes, A. (1998) Computer Based Assessment (Vol 1): a Guide to Good Practice SEED Publications. University of Plymouth.

Chou, C., Moslehpour, M., & Le Huyen, N. T. (2014). Concurrent and Predictive Validity of Computer-adaptive Freshman English Test for College Freshman English in Taiwan. *International Journal of English Language Education*, *2*(1), pp-143. doi:10.5296/ijele.v2i1.4919

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

Chua, Y. P., & Don, Z. M. (2013). Effects of computer-based educational achievement test on test performance and test takers' motivation. *Computers in Human Behavior*, *29*(5), 1889-1895. doi:10.1016/j.chb.2013.03.008

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. doi:10.1111/1467-8535.00294

DeAngelis, S. (2000). Equivalency of computer-based and paper-and-pencil testing. *Journal of Allied Health. 29*(3) 161–164.

DeBoer, G. E., Quellmalz, E. S., Davenport, J. L., Timms, M. J., Herrmann-Abell, C. F., Buckley, B. C., ... & Flanagan, J. C. (2014). Comparing three online testing modalities: Using static, active, and interactive online testing modalities to assess middle school students' understanding of fundamental ideas and use of inquiry skills related to ecosystems. *Journal of Research in Science Teaching*, *51*(4), 523-554. doi:10.1002/tea.21145

Debuse, J. C., & Lawley, M. (2015). Benefits and drawbacks of computer-based assessment and feedback systems: Student and educator perspectives. *British Journal of Educational Technology*. doi:10.1111/bjet.12232

Delen, E., Liew, J., & Willson, V. (2014). Effects of interactivity and instructional scaffolding on learning: Self-regulation in online video-based environments. *Computers & Education. 78*, 312-320. doi: 10.1016/j.compedu.2014.06.018

Dindar, M., Yurdakul, I. K., & Dönmez, F. I. (2013). Multimedia in Test Items: Animated Questions vs. Static Graphics Questions. *Procedia-Social and Behavioral Sciences, 106*, 1876-1882. doi:10.1016/j.sbspro.2013.12.213

García Peñalvo, F. J. (2008). *Advances in E-Learning: Experiences and Methodologies*. Hershey, PA, USA: Information Science Reference (formerly Idea Group Reference).

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis (7th ed.)*. Englewood Cliffs: Prentice Hall.

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. *Procedia-Social and Behavioral Sciences*, *98*, 659-667. doi:10.1016/j.sbspro.2014.03.465

Jawaid, M., Moosa, F. A., Jaleel, F., & Ashraf, J. (2014). Computer Based Assessment (CBA): Perception of residents at Dow University of Health Sciences. *Pak J Med Sci, 30*(4), 688-691. doi:10.12669/pjms.304.5444

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

Kaya, F. & Delen, E. (2014). A computer-based peer nomination form to identify gifted and talented students. *The Australasian Journal of Gifted Education. 23*(2), 29-36.

Kong, X. J., Wise, S. L., Harmes, J. C., & Yang, S. (2006, April). Motivational effects of praise in response-time based feedback: A follow-up study of the effort-monitoring CBT. In *annual meeting of the National Council on Measurement in Education, San Francisco*.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, *67*(4), 606-619.

Lee, M. (2009). CBAs in Korea: Experiences, results and challenges. In F. Scheurermann & J. Björnsson (Eds.), The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing (pp. 187–193). Luxembourg, LU: Office for Official Publications of the European Communities.

Lee, Y. H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assessments in Education*, *2*(1), 1-24. doi:10.1186/s40536-014-0008-1

Mason, B. J., Patry, M., & Berstein, 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. doi:10.2190/9EPM-B14R-XQWT-WVNL

Ogletree, A., Ogletree, S., & Allen, B. (2014). Transition to Online Assessments: A Personal Perspective of Meeting Common Core State Standards in an Elementary School in Georgia. *Georgia Educational Researcher*, *11*(1), 170. Available at: http://digitalcommons.georgiasouthern.edu/gerjournal/vol11/iss1/7

Olea, J., Abad, F. J., Ponsoda, V., Barrada, J. R., & Aguado, D. (2011). eCAT-listening: Design and psychometric properties of a computer-adaptive test on English listening. *Psicothema, 23*(4), 802-807.

Peat, M., & Franklin, S. (2002). Supporting student learning: the use of computer–based formative assessment modules. *British Journal of Educational Technology*, *33*(5), 515-523. doi:10.1111/1467-8535.00288

Pintrich, P. R. (1999). Understanding interference and inhibition processes from a motivational and self-regulated learning perspective: Comments on Dempster and Corkill. *Educational Psychology Review, 11*(2), 105-115. doi:10.1023/A:1022020308002

Prensky, M. (2001). Digital natives, digital immigrants Part 1. *On the Horizon, 9*(5), 1-6. doi:10.1108/10748120110424816

Ricketts, C., & Wilks, S. J. (2002). Improving student performance through computer-based assessment: Insights from recent research. *Assessment & evaluation in higher education*, *27*(5), 475-479. doi:10.1080/0260293022000009348

Rosner, B. (1983). Percentage points for a generalized ESD many-outlier procedure. *Technometrics, 25*(2), 165-172. doi:10.1080/00401706.1983.10487848

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

Schatz, P., & Browndyke, J. (2002). Applications of computer-based neuropsychological assessment. *Journal of Head Trauma Rehabilitation*, *17*(5), 395-410.

Schatz, P., & Putz, B. O. (2006). Cross-validation of measures used for computer-based assessment of concussion. *Applied Neuropsychology*, *13*(3), 151-159. doi:10.1207/s15324826an1303_2

Schatz, P., & Zillmer, E. A. (2003). Computer-based assessment of sports-related concussion. *Applied Neuropsychology*, *10*(1), 42-47. doi:10.1207/S15324826AN1001_6

Sly, L & Rennie L J. (1999) *Computer managed learning as an aid to formative assessment in higher education* in Brown, S., Race, P. and Bull, J. (eds.) Computer Assisted Assessment in Higher Education. London: Kogan Page.

The Joint Information Systems Committee. (2007). *Effective Practice with e-Assessment: An overview of technologies, policies and practice in further and higher education.* Bristol: The Joint Information Systems

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Weinerth, K., Koenig, V., Brunner, M., & Martin, R. (2014). Concept maps: A useful and usable tool for computer-based knowledge assessment? A literature review with a focus on usability. *Computers & Education*, *78*, 201-209. doi:10.1016/j.compedu.2014.06.002

Wirth, J. (2008). Computer-based tests: alternatives for test and item design. In J. Hartig, E. Klieme, & D. Leutner (Eds.), Assessment of competencies in educational contexts (pp. 235–252). Göttingen: Hogrefe.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*, 163-183. doi:10.1207/s15324818ame1802_2

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*, 19-38. doi:10.1111/j.1745-3984.2006.00002.x

Wise, S. L., Bhola, DS, & Yang, S. (2006). Taking the time to improve the validity of low-stakes tests: the effort-monitoring CBT. *Educational Measurement: Issues and Practice, 25*(2), 21–30. doi:10.1111/j.1745-3992.2006.00054.x

Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology, 81*(3), 329-339. doi:10.1037/0022-0663.81.3.329

◈◇◈