Effect of Standardized Electronic Test of Science on Learning Motivation

Osamah (Mohammad Ameen) Aldalalah
Department of Educational Technology
Jadara University, Jordan.

ZyadWaleed Mohamed Ababneh
Ministry of Education, Abu Dhabi, UAE

Mahmoud Mohammad Abu Nawas
Jadara University, Jordan

Reem Ahmad Albatayneh
Ministry of Education, Jordan

Abstract

This study aimed at constructing a science achievement test for the first basic grade, and comparing its application method (paper, electronic) on Its psychometric properties and learning motivation, to achieve that an achievement science test based on the first basic class science textbook of (25) questions was constructed The two versions of the tests were applied on a sample of (103) students in governmental and private schools in Al-Karak Governate (Southern Al-Mazar District) during the first semester 2017/2018. The results of the study showed that there was reliability coefficient for both the tests which is associated with a criterion. There were high validity indicators for both tests. There were statistical differences in the validity in favor of the electronic test whereas there were statistical differences in the reliability in favor of the electronic test. There were statistical differences at the level of statistical significance in transactions trueness which is associated with a criterion, according to the type of test and for the favor of the electronic test. There were statistically significant differences in learning motivation according to type of test and for electronic test.

Keywords: Electronic tests, paper tests, motivation and Psychometric properties.

Introduction

Scienceis one of the most instructional skills that schools focus on. Students begin to learn science formally from the beginning of the lower basic stage; as students depend on this skill in the following stages of study because through it the student will be able to understanding the various textbooks of the curriculum. On the other hand, science plays a role in the development of the student's personality and enables him to be familiar with different knowledge (Aduhaini, 2017). Science is very important in the school life of the students. It is the basis of logical sciences, and a key to acquiring many subjects of study.

Modern education is concerned that the sciences materials offered to students in the early years of study are easy, simple and far from artificiality, so as to suit the time and mental age of the students so that he can deal with them with passion and desire (El-Borai, 2013). The successful teacher is the one who creates the appropriate environment for the student to acquire different experiences during the learning process; and no doubt that the teacher can achieve this through the preparation of a good teaching program working on the development of scientific concepts in the student with the need to include the basic skills of thinking such as recognition and understanding of the meaning and knowledge of the main ideas of what he learn (Shakhriti, 2009). There are many tests that measure students' skills, which help the teacher evaluate the student continuously, contribute to the brainstorming, and help the teacher to identify educational problems that some students may experience (Wise, sevcik , Morris, Lovett, Wolf, Kuhn, Meisinger & Schwanenflugel, 2010).Tests are one of the most important tools of evolution. They are one of the main indispensable tools in the teaching/learning process due to their ability to measure the level of achievement of individuals and to recognize the extent of realizing the curriculum for which the goals were designed, as well as their ability to identify points of weaknesses and strength of the students and the progress achieved by the educational institution so that it can improve and develop the educational process and take it to the maximum possible success and excellence and move forward for the better (Sarayra, 2011). Najjar (2010) points out that test are one of the most important methods that have been used by humans since ancient times.

Humans began to test different things around them to identify their nature, and then used these tests in the teaching-learning process to learn about the effectiveness of this education, students' achievement level, the achievement of the objectives, educational outputs, and the teacher's various educational activities that help raise the skills of the achievement of learners.Test is defined as "a series or set of questions or tasks that the learner is required to respond to verbally, in writing, or in performance (scientifically). Further, the test should include a representative sample of all possible questions and tasks related to the property measured"(Ibrahim and Abu Zeid, 2010, p. 520).Najjar (2010) points out that the achievement test is a set of stimuli (verbal or written questions, pictures or drawings) prepared to measure a behavior quantitatively. The test gives a score, value or rank to examinees. Whereas Ibrahim and Abu Zeid (2010) point out that the achievement test "is the one designed to assess students' knowledge and skills that they have learned or trained on. "Oudah (2010) adds that achievement tests are designed to measure the extent to which the learner has acquired knowledge and skills in one of the different educational areas at the end of a particular study period, in which the learner answers a sample of the questions that represent the content of the educational material. School tests can cause anxiety and panic to some learners. Learners may reach the extent of nervous breakdown and that may affect the results of the exams because of the learner's sense of distrust and helplessness because the test questions often measure some of the goals that the student has not trained on. Furthermore, tests measure the difficult parts of the study content, as well as they discourage creativity and focus on the minimum thinking skills such as remembering, understanding and comprehension (Murad & Sulaiman, 2002).

Apart from traditional tests, computerized tests are one of the techniques that can be employed to overcome the negatives that accompany paper and pen tests, or employ them to provide other channels to increase achievement and retain information among learners. The test is finalized by random selection from a wide range of questions in the Bank of Questions, ensuring proper representation of the full test dimensions according to the specification table (HRD, 2010). Al Ghareeb (2009) defines e-learning evaluation as "a process of employment in information networks, computers, educational software, and multi-source learning material, using assessment tools to collect and analyze student responses, helping teachers to discuss and determine the impacts of the programs to reach a quantitative judgment based on quantitative or qualitative data that relate to academic achievement". Ahmed (2012, p. 191) points out that electronic tests are a computer application that can be employed to overcome some of the difficulties that can hinder the implementation of traditional (paper) tests; or employ them to provide other channels in order to increase students' educational achievement and consolidate information and develop self-learning skills. Electronic tests are an easy way to evaluate the student electronically as they enable the teacher to prepare them easily; apply them to the students then they become electronic and immediate tests; thus ensuring the credibility and transparency of the scoring process (Alomary & Iyadat, 2016). Abdelhamid (2005) defines electronic testing as a continuous and structured educational process aimed at evaluating student performance through the use of electronic networks .Summadi (2009) summarizes the role of the computer in tests by writing questions, constructing and storing, applying, scoring, analyzing, and obtaining a comprehensive report of the results of the test. Hence, Al-Absi (2010) emphasizes that the importance of computerized tests stems from being a tool that helps in evaluating learners and determining the extent to which educational goals have been achieved. The most important things that show the importance of these tests is being able to identify the strengths and weaknesses of the learners; measure their achievement and progress; stimulate their learning; evaluate the teaching methods used; evaluate the curriculum and its relevance to the needs of learners; provide parents and decision-makers with feedback on the level of achievement of their children, and evaluate the educational program as a whole (Badawi, 2014). Yurdabakan (2012) believes that electronic tests will address errors that may occur in manual tests; remove all human errors that may occur in paper tests and give the student an opportunity to get the test result immediately.As for the types of tests, it is found that specialists in the fields of education in general, and in measurement and evaluation in particular emphasize that tests and measurements that are designed and applied in certain environments may not be valid and suitable for application in other environments, even if similar in some circumstances and variables, as there are effects that make it necessary to ascertain the appropriateness of these tests to the environment in which the phenomenon to be measured; therefore, there is a clear interest among researchers and specialists in these areas of legalization and development of measures and tests used in the study the psychometric characteristics to fit the environments in which they work. Abu Kassarah and Ziad (2015) see a need to pay attention to the psychometric characteristics of the tests with a focus on validity, reliability as well as on the difficulty and discrimination coefficients to indicate the psychometric characteristics of the test. Dweidri (2000) thinks that when preparing the tests the researcher must be concerned with their validity, reliability and objectivity. Validity is the most important condition of a good test, measuring what is to be measured (Saber and Khafajah, 2002). Validity is a feature and characteristic of a tool that is linked to the primary purpose for which the measure will be used, and the decision to be made based on the results of this measure (Ghamdi, 2003). There are several types of validity that must be available in each test as indicated by Al Assaf, 2011; Dweidri, 2000; Saber and Khafajah, 2002.

First: Validity of Content which indicates the degree of the representation of the test items of the content to be measured. Among its types is the virtual validity which is interested in the content of the items and judges the extent of conformity to the content of the study material; and face validity when the measure represents a sample of the behavior to be measured. This requires that experts arbitrate the items of the test, and assess the extent of measurement as it seems clear for the attribute that the items are prepared for its measurement. Second: criterion-related validity: It is the extent to which the performance in the test items we want to do at this time, and the extent to which it corresponds to the items of another test that its validity has been proved in a short period of time. Here, the semantics of this validity can be verified through the correlation coefficient. We can distinguish between two types of validity of the test: predictive validity which isthe accuracy of the test prediction means the future behavior shown by the performance of the individual at stake; and predictive validity, which is determined by comparing the results of the test whose validity is sought with the results of the measure another was applied at the time of application to test or shortly after. The validity of building (concept) (Construct Validity) is to derive hypotheses about the results of the test, and to verify the hypotheses logically and experimentally.

As for the reliability of the test: the test is reliable if it can give the same results if repeated regardless of the circumstances surrounding it (Assaf, 2011). Darwaza (2001) considers the reliability of the test one of the qualities that must be met in a good measuring instrument. Saber and Khafajah (2002) point out that when the same test is repeated, stable results are given. In the case of ordering the examinee in his group, this order does not change when the test is re-applied under the same conditions. There are four types of reliability: First reliability coefficient: the reliability coefficient is obtained in this method by applying a test to a group of people and then reapplying the same test on the same group after a period of time and calculating the correlation coefficient between the scores of the group members on that test of the two groups (Nabhan, 2004). Second: Equivalence reliability is calculated by giving two equal forms of content, variances and averages for a given test for the same group at the same time. The correlation coefficient is calculated after a short period, so that these tests have the same difficulty and discrimination of the items and the type and length of the tool. Application procedures should be unified in both times in terms of response time, correction and instructions (Al-Nabhan, 2004). Third: reliability and equivalence reliability is calculated by combining the methods of reliability and equivalence by applying one of the two forms, so that the time of application of the second form in this method is long, and by calculating Pearson correlation coefficient of degrees in the two forms (Allam, 2006). Fourth: reliability of internal consistency: by estimating reliability using one test and applying it once (Allam, 2006).

Motivation

Inspiration a gathering for motivates spurs those scholar on fulfill targets (Al-Jarrah et al. , 2014). This situation as standard impacts around unique behavior as it manufactures that rationality of the behavior should fulfill those obliged targets (Hadeh, 2013). The inspiration enacts Furthermore coordinates mankind's direct (Negovan Also Bogdan, 2013). Various components effect independent practices, for example, the ability should assume out those direct and the approachability about fitting conditions, in any case of if such states would accessible, it doesn't guarantee those innovation of the conduct, instead it depends upon the size of an individual's inspiration (Salem, Kabylie, &Khalefah, 2012).To investigators What's more instructors' indistinguishable inspiration need been a standout amongst the way plans used to elucidate different extents of execution. It intimates will elucidate contrasts in the measure from claiming effort associated with Taking in errands Also may be in this way anticipated on a chance to be solidly recognized for contrasts done extents from claiming execution. During its easiest, inspiration need been identified with the measure from claiming insightful vitality typically used over Taking in exercises, and this incited An conviction that inspiration Might been seen as a unfaltering typical to the person, for a standard for character (Aldalalah, Eyadat&Ababneh, 2015).

Problem of the study and questions:

The problem of the study is shown by examination of the studies (Al-Awasa, 2016; Rawashdeh, 2016; Al-Amr, 2015; Al-Mujdah, 2014) which indicate that although it is important and necessary to design and build appropriate tests for the local environment that are originally built according to the cognitive and cultural standards and values prevailing in this environment; however, these tests with psychometric characteristics may not be suitable for application to the target sample for the lower elementary stage in Karak governorate. Hence, by informing the researchers that there are no computerized tests in science for the first grade students in the absence of adequate computers in some schools, and that most of the Ministry of Education's focus was on the traditional tests despite the lack of psychometric characteristics and the characteristics of good testing. In order to achieve a good level of science, a well-designed test is required to measure the level of science among students, so that the students' scores contribute to the students' future prediction and to assess their science strengths and weaknesses.

Specifically, this study answered the following questions:

- What are the significant of validity (internal consistency validity, content validity and validity associated with the criterion) that is available for the achievement test in science for first grade students?
- What are the significant of reliability (internal consistency Kronbach Alpha) that is available for the achievement test in science for first grade students?
- Is there a difference in the psychometric characteristics of the science achievement test due to the different method of applying the test (paper, electronic)?
- Is there a difference in the learning motivation due to the different method of applying the test (paper, electronic)?

***Previous studies*:**

This part of the study will review previous studies according to the study variables, as follows:

A. Studies of the difference between the two types of testing:

Al-Amri's study (2007) aimed to reveal the comparison between the computer-based test and the paper test and its effect on the achievement of the learners. A sample of 167 Saudi students in the medical field was selected. The researchers used various tools for collecting the data: the study found that there were no statistically significant differences in achievement and all study variables when using the computer-based test, and based on the results of the study. The researchers suggested that the paper test be used instead of the computer-based test.

Sim& Horton (2005) conducted a study that was applied to 20 students in the third grade in the UK. The aim was to identify the differences in the students' performance in the paper and computer test, and to identify the students' attitudes toward computerized testing. The study showed that 50% of the students performed better in the paper test while 25% performed better in the computerized test. 25% of the students performed the same in paper and electronic tests. The study also showed preference for computerized tests on paper tests.

In a study conducted by Clariana& Wallace (2002) and applied on 105 undergraduate students in business administration to study the difference in 4 factors between paper and electronic tests. These factors are knowledge of content, computer literacy, competitiveness, and gender. Students were exposed to two copies of the same test: one paper and the other computer. The results showed superiority for students who were tested on the computerized one. There was no difference in the factors related to gender, competitiveness and computer knowledge; while there was a difference in the knowledge of content.

Kingston (2009) conducted a comparative study between computer testing and paper and pen testing in the science subject. The study examined the results of 81 studies completed between 1997 and 2007; the expected effect size across all studies was very small. The methods of factor analysis were used to verify whether the variable of grade (primary, intermediate or high) English grammar, mathematics, reading, science and social studies, have an impact on comparison. The findings of the study showed that there were no statistically significant differences in the class variable, while differences were found on the subject variable in favor of English language literature followed by social studies and mathematics.

Deangelis (2000) conducted a study with the aim of familiarization with the extent of equivalence of computerized tests, paper and pen tests, and students' attitudes and perceptions towards them. The researcher chose a sample of 30 students from the first year in the field of dentistry randomly; and then they were divided into two groups. The first group took the paper and pen test while the second took the same test computerized. After a while, the two groups were switched. The first group took the computerized test and the second group took the paper and the pen; and then a questionnaire was distributed to the students to measure their attitudes and perceptions towards computerized tests. The findings showed that the students' achievement in the computerized tests was better than that of the paper and pen tests, with a statistical indication in favor of the computerized test. Further, the findings showed that the students' acceptance of this type of tests was between intermediate and high as computerized tests save time and effort, give students a faster answers, and provide quick feedback.

In a study by Wang, Jiao, Young, Brooks and Olson, 2008, under the title "A Comparison between the Use of Computer-Based Tests and Paper Tests and Their Impact on Learners' Performance in Reading Assessment: a Comparative Study of the Impact of the Type of Test" the researchers selected a sample of 22 studies from previous studies on this subject conducted between 1993 and 2005. They analyzed these studies and concluded that there were no statistically significant differences between computerized tests and paper and pen tests on all study variables: test design, sample size, and computer skills. In a study by Kapoor & Welch (2011) in Texas, titled "A Comparison of Test Pencil Paper (TPP) and Test Based Computer (TBC) in Mathematics Assessment, and the Effect of the Test Management on Them", the two researchers used a sample of (689) male and female students in the fifth primary grade and 676 students in the eighth grade.

The study concluded that the analyzes conducted at grade levels indicated that the fifth graders found that paper and pen tests are easier than computerized testing and that the eighth graders found that computerized testing was easier and that there was no impact of the way in which the test was administered.

The study of Khazzi and Zakri (2011), which aimed to test the equivalence between electronic and paper tests in measuring university achievement and the impact of students' exposure to electronic tests on their attitudes towards them, the experimental method was used, where 316 students at the Faculty of Education at Kuwait University were given two identical copies of the tests: paper and electronic. This was accompanied by measuring the students' attitudes using a questionnaire about electronic tests before and after exposure. The findings showed the equivalence of electronic and paper tests in the measurement of students' academic achievement, with statically significant differences in the time required to perform the test in favor of electronic testing. Further, the findings of the study showed high attitudes of students towards electronic tests because of exposure to them. The study recommended adopting the use of electronic tests in university education in similar educational and humanitarian disciplines as well as conducting similar studies.

B. Studies on the psychometric characteristics of tests:

This section will deal with some studies conducted on the psychometric characteristics of items and tests. The studies were arranged from oldest to newest.

Al-Kahlout (2002) refers in his study aimed at comparing the psychometric characteristics of both multiple choice tests and complementary tests. The measurement consisted of 23 items for each type. The sample of the study was selected from the UNRWA schools in Jordan, where the sample number was 451 students from the sixth grade. The findings of the study indicated that the coefficient of reliability of the supplementation test calculated in Cronbach Alpha method and the method of repetition was greater than the reliability coefficient of the multiple choice test. Moreover, the findings indicated that the mean difficulty coefficients for the multiple choice test were greater than the mean of the difficulty coefficients for the complement test, and that the mean of the discrimination coefficients for the supplementation test was greater than the that of the mean of the discrimination coefficients of the multiple choice test.

Yassin's (2004) study aimed at estimating the psychometric characteristics of the criterion-referenced test in chemistry for the first scientific secondary grade according to the classical and modern theories of measurement. The sample of the study consisted of 481 male and female students distributed in 14 sections. The sample was selected by the random cluster method. A criterion referenced test in chemistry made up of 52 items was applied on the sample of the study. This study concluded that the estimation of the psychometric characteristics of the test (validity and reliability) was accomplished according to the classical theory; where the coefficient of validity in terms of the criterion was (0.84), the reliability coefficient of Kronbach Alpha was (0.90); the reliability estimate according to the modern theory was accomplished by using the Rush form; the reliability coefficient of the test was (0.99), and the reliability coefficient for the individuals was (0.88).

The study of Tarrant & Ware (2010) compared the psychometric characteristics of multiple choice tests with three or four alternatives used in the assessment of nursing students. In order to achieve the objectives of the study, the researchers applied the multiple choice test with four alternatives to a survey sample to examine and compare the psychometric characteristics of the test items. Using the analysis of the items, the weak equation was identified in the answer process. The researchers prepared the final form of the test, so that the test consisted of 41 items in each form, the first form contained three alternatives while the second contained four alternatives. The results of the study showed that tests containing three alternatives were more effective, despite the lack of disguises, because of the strength of these disguises. The results of the study indicated that the disguises used in the multi-choice test became highly discriminate when the omitted disguises were not frequently selected in the answering process.

The Methodology of the Study

The descriptive method was used as a way to build the achievement test in science by analyzing the content of the science book for the first grade and the general framework of the science curriculum for the first grade to extract the content elements and cognitive levels. The science test was built to measure the cognitive level of the students. The quasi-experimental method was followed to compare the effect of the method of applying the test on its psychometric properties.

The Study Population and Sample

The study population is composed of all the first grade students in the public and private schools under the Ministry of Education in the governorate of Karak (Southern Mazar) and registered for the first semester of the academic year (2017/2018) in the second grade (1814). The sample of the study was chosen according to the intentional method.

In the first stage, 2 schools were selected. One class was chosen randomly from each school, in which the electronic test was applied. , And 2 other schools were selected randomly. One class was randomly selected from each school for the paper test. The study sample was 103 students.

Study Instruments

To achieve the objectives of the study, a multi-type test in science was constructed for students, who completed the first grade, So that the test includes two types in the method of application (paper, electronic). It should be noted here that the electronic test differed from the paper test because it has several specifications: That the electronic test incorporates the colors, and the movement that is in the paragraph makes it easier for the student to understand the required question, since it does not need to read the question, and each paragraph in the test one answer only correct, The process of building this test has gone through the following stages:

- Four courses were selected in the science lessons to build a paper and electronic test covering all of these lessons.
- The teaching objectives of these lessons, which reached (12) teaching objectives, as stated in the textbook, After reviewing the science curriculum, A number of educational supervisors have been employed as arbitrators, and these teaching objectives have been distributed among the lessons.
- The team was asked to classify the 12 teaching objectives in five levels of knowledge: memory, understanding .And then distribute the teaching objectives and their relative weights according to the levels of knowledge according to the lessons. Table (1) shows this.

Table (1).Table of test specifications according to educational objectives.

Lessons	Percentage	**Classification of teaching objectives**			
		memory		**understanding**	
		number	**Percentage**	**number**	**Percentage**
1	25%	2	66.7%	1	33.3%
2	25%	2	66.7%	1	33.3%
3	25%	2	66.7%	1	33.3%
4	25%	1	33.3%	2	66.7%
Distribution of the number of test paragraphs		7paragraphs		5paragraphs	

- The test paragraphs, the number of paragraphs (12) By helping from first-class primary teachers to researchers, And then presented to the arbitrators to express their opinion and observations on them in order to validate the content of the test, two questions were deleted to duplicate their content, one question was deleted for a long time in the solution, and one question were deleted because they did not belong to the content of the teaching material. Two questions were combined with on equation. The number of paragraphs was 30, One score for each test paragraph was approved in the case of the correct answer and zero if the answer was incorrect or not answered.
- The pilot study sample consisted of 49students who were selected in the available method, In order to estimate the difficulty and discrimination coefficients of the test paragraphs, Tables (2 and 3) show the difficulty coefficients and the discrimination coefficients for each of the test paragraphs.

Table (3): Difficulty and Discrimination Transactions for each item of the paper test

paragraphs number	Transactions Difficulty	Transactions Discrimination
Q1	0.34	0.43
Q2	0.41	0.53
Q3	0.38	0.50
Q4	0.51	0.62
Q5	0.51	0.56
Q6	0.48	0.53
Q7	0.48	0.50
Q8	0.43	0.35
Q9	0.62	0.53
Q10	0.67	0.56
Q11	0.52	0.50
Q12	0.40	0.43

Table (2): Difficulty and Discrimination Transactions for each item of the electronic test

paragraphs number	Transactions Difficulty	Transactions Discrimination
Q1	0.37	0.40
Q2	0.44	0.51
Q3	0.61	0.40
Q4	0.63	0.59
Q5	0.51	0.59
Q6	0.46	0.40
Q7	0.43	0.44
Q8	0.42	0.59
Q9	0.44	0.40
Q10	0.44	0.44
Q11	0.61	0.48
Q12	0.61	0.66

Table (2) shows that the difficulty coefficients of the vertebrae of the paper test range from 0.34 to 0.67 and the discrimination coefficients ranging from 0.35 to 0.62. Based on the acceptable range of difficulty and discrimination of paragraph.So that the maximum mark for the paper and electronic final test (12) is shown in Table 3. Table 3 shows that the difficulty coefficients of the paragraphs for the electronic test range from 0.37 to 0.61 and the discrimination coefficients range between 0.40-0.66.

Motivation questionnaire

This questionnaire is used to measure the learning motivation of first grade primary students. The questionnaire of Motivation formed of notification asking about feelings. Generally, the questionnaire consists of 10 items. These items were rated using a 3 point Likert scale with the following anchors: 1 = = Agree; 2 = Neutral; and 3 = Disagree. This questionnaire is adopted and adapted from Aldalalah, Eyadat&Ababneh (2015). The pilot study consisted of 49 participants. The researchers used Test-Retest to check the reliability of the instrument. That unwavering quality coefficient for this instrument flying (The Arabic version) might have been registered by that execution about Cronbach alpha whereby it might have been 0. 79 to those entirety scale. That interior consistency in this instrument flying (Arabic version) might have been 0. 82.

Study variables:

1. Independent variables: Method of applying the test (paper and electronic).
2. Dependent variables: Psychometric characteristics have two categories (validity, reliability).And learning motivation.

Result of the study:

The aim of this study was to construct an achievement test in science for the first grade, and to compare the effect of its method of application (paper, electronic) on its psychometric characteristics and motivation. The main findings of the study are presented below:

- **Findings of the first question**: What are the indications of validity (internal validity, validity of content and criterion-related validity available for the Science test for first grade students in both paper and electronic forms?

First: Internal validity:

To estimate the internal validity of the test, the corrected correlation coefficient for each type of test was calculated and differences were found between them as in Table 4.

Table (4): internal consistency validity of the test (paper and electronic)

Paragraphs	Electronic	Paper
1	. 460**	.349**
2	. 476**	.284**
3	. 558**	.366**
4	. 553**	.401**
5	.434**	.391**
6	. 504**	.305**
7	. 502**	.510**
8	. 510**	.572**
9	. 409**	.525**
10	. 467**	.510**
11	.552**	.449**
12	. 412**	.501**

It is noted from Table 4 that the coefficients of the internal validity of the electronic test ranged between 0.409-0.558, and the paper test ranged between 0.284 -0.572.

Further, the findings of the study indicate the availability of indicators of validity for the two tests. This may be explained on the basis of the test questions which were prepared based on the theoretical literature on the subject, in addition to employing the expertise of the researchers combined with the opinions of the arbitrators.This helped the researchers construct test items with scientific standards.

Content Validity:

This type of validity has been achieved through the procedures used by the researchers in the design of the test, as well as ensuring the relevance of the test subjects to the content of the material to be tested, verifying the relevance of these

items to the expected educational objectives to be accomplished, and verifying the correctness of the language of the items and their relevance to the level of the students.

After preparing the test in its initial form, it was presented to a group of arbitrators who were a group of supervisors and teachers of the Arabic language in the Ministry of Education, and some of the teachers of the subject and specialists in this area to identify the appropriateness of the test for the purpose that it was prepared for, and the relation of its items to the learning of science among the first basic grade student in Karak governorate, as well as to be familiar with the opinions of the arbitrators on the correctness and clarity of the wording of the test items. Some of the test items have been modified in the light of their opinions, as mentioned below.

This type of validity was achieved for the test through the procedures used by the researchers in designing the test. The researchers wrote a large number of items in the initial phase that amounted to 40, and then they were presented to the teachers and supervisors of the first grades to express their opinions and observations on the test items. They differed on 5 questions because of the replication of their content and the educational goals they measured. Further, they differed on one question as it required a long time in the solution; and on other two questions because they did not belong to the content of the teaching material. Moreover, three questions were combined into two as belonging to the same property; therefore, the number of items of the final test was 25. On the other hand, the members of the arbitration team agreed on all the final test items, and the percentage of difference was very small compared to the proportion of agreement among them, which indicates the achievement of virtual and face validity.

Criterion-related Validity:

The indicators of the criterion validity of the test were obtained by calculating the Pearson correlation coefficient between the total scores of the individuals on the items of the test forms and the total scores of the individuals at the end of the second semester assuming that they have an acceptable degree of validity and stability. Therefore, in coordination with the schools in the Directorate of Education for the Southern Mazar Region, the researchers were keen to make a unified test at the end of the second semester for science for the sample members of the study and in line with the Ministry of Education's keenness that the tests applied should be unified for the first grade students as in Table 5.

Table (5): Criterion-related Validity

Test	Validity coefficient *
paper	0.701
electronic	0.687

The sign (*) indicates differences with statistical significance at the significance level α=0.05

It is noted from Table 6 that the coefficient of the criterion-related validity of the paper test was 0.568 and was less than the electronic test which reached 0.675. The difference between the coefficient of validity with statistical significance at the level of significance was α=0.05. The correlation coefficients between the students' performance on the science test prepared by the researchers (electronic and paper) and the final score by the teacher were positive and high. This means that the ratio of the common variance between the scores on each test and the teacher scores was high, possibly because of the similarities between the two tests.

In this study, the findings of electronic tests are higher than those of the paper ones. This result can be attributed to a higher variance between the scores of each individual on the electronic test than that in the paper one. This is because the criterion was conducted in paper and the students were less accustomed to computerized tests.

- **Findings of the second question:** What are the significances of reliability (reliability of internal consistency Kronbach Alpha) available for the test of achievement in science for the first grade students?

To answer the question, the reliability coefficient for the two test forms was estimated using the Alpha Kronbach equation; and Table 6 shows these findings:

Table (6): reliability coefficients of internal consistency

Test	reliability
paper	0.684
electronic	0.851

It is noted from Table 6 that the reliability coefficient of the electronic test was 0.851 and was higher than that for the paper test, which was (0.684).

The findings indicated that the coefficient of test reliability estimated by the Cronbach alpha formula for the electronic test form was more stable than that in the paper tests. It was 0.851 for the electronic test and 0.684 for the paper test.

This can be explained by the fact that electronic tests have a number of characteristics, including: being interactive and flexible;with accurate results; safer; conducted and scored easily and immediately; having impartiality that may result by scorers while scoring them to test a particular class of students. Further, they create a kind of challenge between them and the student where the student gets an immediate result of his performance, which increases his interest. Moreover, the student is less busy and wasting time by erasing and adjusting errors; pen sharpening and the like, in addition to preoccupation with those around him. Therefore, he finds in the computer a motivation factor; this can also be explained by the fact that today's students are considered "digital instinctive" because of engaging in electronic and educational games on computers, telephones and other modern means that attract them. The student finds here a bit of a challenge to answer questions.

- **Findings of the third question:** Do the psychometric characteristics of the science test among the first grade students differ according to the method of applying the test (paper, electronic)?

First: **Regarding the validity of internal validity**:

In order to provide the indications of the validity of the internal validity of the test, the corrected correlation coefficient for each type of the test was calculated and differences were found between them as in t-test Table 7.

Table (7): Means, standard deviations and t-test for two independent samples

	test	N	Mean	Std. Deviation	t	Sig
Test	paper	12	.492	.047	2.577	.017
	electronic	12	.418	.087		

The results of Table (7) showed statistically significant differences at the level of significance ($\alpha = 0.05$) between the mean of the degree of internal validity due to the variable type of test (paper, electronic), Where the level of statistical significance .017

This is based on the test questions, which were prepared based on the theoretical literature on the subject, as well as the use of the expertise of the researchers, which helped them build items with scientific standards in the writing of the items of the test, and consultations from specialists and colleagues, as well as control procedures. Through acquaintance with previous studies, the researchers sought to find equivalence in the electronic and paper tests in the science achievement test. Because e-tests help teachers evaluate the degree of student retrieval and participation, they can test their ability to participate in all student postings in mailing groups, panel discussions, and save results for tests and assignments in student record databases as easily accessible at all times and times.

Second: reliability:

To test the significance of differences in reliability coefficients between the paper and electronic tests,

Table (8): reliability coefficients of internal consistency

Test	reliability
paper	0.684
electronic	0.851

This may be due to the fact that the electronic tests have characteristics such as interactivity, flexibility, accuracy of results, safer, easy to carry out, immediate correction, and the impartiality that may be caused by the scorers during their scoring for a particular class of students. Additionally, the scores may be influenced by scorer's psychological status and temperament when scoring the paper tests, the accuracy may increase and that is reflected in the assessment of students. Moreover, paper scoring may differ from a scorer to another. However, in the computerized test there is no place for this, and this type of tests increases the motivation of the student as each new question is considered a new motive or stimulus that has to be solved as a challenge.

Furthermore, the student in the computerized test is fully aware that the teacher does not put the score and it is put objectively and electronically, which generates the perseverance and motivation to understanding and persistence. The electronic tests do not require scorers. The student takes the result immediately, which reduces the errors of scoring and the difference among scorers, i.e., scorer's stability. Hence, there is no need for the reviewers to perform the review and audit, which in turn plays a role in that the electronic tests are more stable than the paper tests.

Findings of the forth question is there a difference in the learning motivation due to the different method of applying the test (paper, electronic)?

Means and standard deviations of student motivation were calculated and the t-test of two independent samples was applied to determine the difference, table (9).

Table (9): Means, standard deviations and t-test for two independent samples

	test	N	Mean	Std. Deviation	t	Sig
Motivation	paper	200	3.12	1.12	2.183-	.030
	electronic	200	3.37	1.16		

The results of Table (9) showed statistically significant differences at the level of significance ($\alpha = 0.05$) between the arithmetic mean of the degree of motivation due to the variable type of test (paper, electronic), Where the level of statistical significance .030

The exhibit examine discovered that those E-test system learners benefited more than their counterparts who have the same qualities in any case were utilizing paper test clinched alongside inspiration. E-test assumes a noteworthy part for upgrading inspiration taking in for understudies. The E-test depended on the collaboration the middle of two channels. For A deliberate mental activity, the content that is shown the ai route fortified deliberate supposing what's more connection. The understudies Might thereby recover those information All the more effectively. On the other hand, understudies were over need of a instructing perusing that responds to their distinctive contrasts Furthermore transforming capacity from claiming people. Therefore, learners were over requiring of a instructing perusing that acknowledged their qualities. The E-test system might have been supportive for such person as they remember and seeing content for new learning being incorporated under their past information.

Recommendations:

In light of the findings of the study, the following can be recommended:

- Providing an electronic test benefits teachers in measuring the levels of student achievement in all subject
- Encouraging teachers by the Ministry of Education to use electronic tests for their reliability and the need to train teachers in their construction
- Re-applying the measurement to other categories of the same stage on which the test was applied and getting familiarized with its validity and reliability, and deriving its criterions
- Educating students in all age groups about the importance and effectiveness of computerized tests
- Working on the development of computerized tests to give at the end of the test a detailed disclosure of the student that includes the correct answers and the wrong ones.

References

Abdel Hamid, Mohamed Abdel Hamid, (2005). Learning system across networks. Cairo. World books.

Abssi, Mohammed. (2010). Realistic assessment in the teaching process. Amman: Dar Al Masirah for Publishing, Distribution and Printing.

Abu Kassara, Mansour. Ziad, Rashid, (2015). The psychometric characteristics of the Algerian version of the scale of the expectations of the general self-efficacy of secondary school students. The survey of psychological and educational sciences, University of Oran: Algeria.

Ahmed, (2010).Measurement and evaluation in the teaching process. Irbid: Dar Al Amal for Publishing and Distribution.

Ahmed, Yasser, (2012). The use of computers in education. Dar Al-Zahra: Riyadh.

Aladdro, Lair, (2015). The impact of educational software in providing reading skills for kindergarten children. Unpublished Master Thesis, Mu'tah University.

Al-Amri, S. (2007).Computer-based vs. Paper-based Testing: Does the test administration mode matter? The BAAL Conference (2007), 101-110. Retrieved March 18, 2017, From :www.baal.org.uk/proco7/33_saad_al_amri.pdf

Al-Borai, Abdullah, (2013).The effectiveness of a computerized program for the treatment of learning difficulties in the third grade students. Unpublished Master Thesis, Islamic University: Gaza.

Al-Dahini, Rasha, (2017).The lack of reading and behavioral indicators characteristic of the third grade students.Unpublished Master Thesis, Islamic University of Gaza.

Aldalalah, O., M., A., Eyadat, Y., A. &Ababneh, Z., W., M. (2015).Effects of webquest on the achievement and motivation of Jordanian University students of (independent & dependent) cognitive style.World Journal on Educational Technology. 7(2), 119-135.

Al-Ghamdi, Said Hassan, (2003). The extent of the different characteristics of the instrument of measurement in light of the variance of the number of alternatives to the response and the stage of the study Master Thesis, University of Umm Al-Qura, Makkah.

Algharib, Zaher Ismail, (2009). Electronic courses, design, production, publishing, application, and evaluation. World Books: Cairo.

Alkhizyu, Fahd and Zacari, Mohammed. (2011). Equivalence of electronic tests with paper tests in the measurement of academic achievement: A pilot study on students of the Faculty of Education, Kuwait University. Gulf Studies and Arab Countries, 143: 1-32.

Allam, Salahuddin Mahmoud, (2006). Measurement and evaluation of educational and psychological foundations and contemporary applications. Cairo, Arab Thought House.

Al-Mawajdah, Huda, (2014) Difficulties in Teaching Arabic Language in the Lower Basic Stage in the Southern Mazar District from the Point of View of the Teachers of the Article. Unpublished Master Thesis, Mu'tah University.

Al-Nabhan, Musa, (2004).Principles of Measurement in Behavioral Sciences, 1, Amman, Dar Al Shorouk Publishing and Distribution.

Al-Sarayra, AyatJaafar (2011). An analytical study of the patterns of achievement test questions among the social and national education teachers of the fourth and fifth grades in the Southern Mazar district. Master thesis unpublished in general curricula and methods, Mutah University.

Al-Smadi, Ezzat, (2009). Computerized tests and question banks. A working paper presented in the training program for faculty members at Umm Al-Qura University, July 7,

Assaf, Saleh, (2011). Introduction to research in behavioral sciences. I. Riyadh: Dar Al-Zahra.

Awasa, Du'aa, (2016). The Effect of Designing Interactive Educational Content Using Articulate Storyline on Developing Reading Skills among Kindergarten Students in Al-Mazar Al-Janabi Schools Unpublished Master Thesis, Mu'tah University.

Badawi, Mohamed, (2014) Effectiveness of a proposed program in e-learning to develop the skills of designing electronic tests and the trend toward electronic evaluation among postgraduate students. International Specialized Educational Journal, 3 (5): 146-176.

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

Darwaza, AfnanNazir, (2001). Educational questions and school assessment. I 3, Dar Al Shorouk for Publishing and Distribution: Amman.

Dawidri, Raja Wahid, (2000). Scientific research bases its theory and practice. Dar Al-Fikr: Damascus.

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

Ibrahim, Mohammed and Abu Zeid, Abdul Baki, (2010).Educational Research Skills. I 2, Dar Al Fikr for Publishing and Distribution: Amman.

Kahlout, Ahmed Ismail, (2002). Comparison of the psychometric properties of both multiple choice tests and complementary tests.Journal of the Center for Educational Research, University of Qatar. 11: (22) 127-153.

Kapoor, S., & Welch, C. (2011).Comparability of paper and computer administrations in terms of proficiency interpretations.A paper presented at the annual meeting of the National Council on Measurement in Education New Orleans, LA:1-17. April 2011.

Kingston, N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations.A synthesis.AppliedMeasurement in Education, 22(1), 22-37.

Murad, Salah Ahmad and Sulaiman, Amin Ali, (2002). Tests and standards in the psychological and educational sciences. Modern Book House: Cairo, Egypt.

Mustafa, Mansouri and Wafa, Ben Arum. (2016). Reading difficulties for second and third year students. Psychological and Educational Studies, 15, 17-31.

Najjar, Nabil, (2010). Measurement and Evaluation, Applied Perspective with Software Applications SPSS Dar Al Hamed Publishing and Distribution: Amman.

Omari, Mohammed, Iyadat, Yusuf, (2016). The perceptions of faculty members and students on computerized tests in the process.Learning instruction, Journal of Educational Sciences, 12 (4), 478 - 469.

Rawashdeh, Dreams, (2016). Educational problems facing kindergartens from the point of view of female teachers in the southern Jordan Valley. Unpublished Master Thesis, Mu'tah University.

Saber, Fatima and Khafaja, Mervat, (2002).Foundations and principles of scientific research. 1, Radiology Library and Printing Press: Alexandria.

Shakriti, Sawsan, (2009). The impact of a proposed program in the development of some reading skills among the third grade pupils in UNRWA schools in northern Gaza. Unpublished MA thesis, Islamic University, Gaza.

Sim, G. & Horton, M. (2005).Performance and attitude of children in computer based versus paper based testing. [Online]. Available at: http://www.uclan.ac.uk/facs/destech/compute/staff/read/Publish/ChiCi/references /performance_and_attitude.pdf

Tarrant, M. & Ware, J., (2010). A comparison of the psychometric of three and four- option multiple choice questions in nursing assessment. Nurse education today, 30,539-543.

Training and Human Development Unit, (2000).User Manual for Electronic Testing System.

Wang, S., Jiao, H., Young, M., Brooks, T. & Olson, J. (2008).Comparability of computer-based and paper -and-pencil testing in K–12 reading assessments:A meta-Analysis of testing mode effects.Educational and psychological measurement, 1(68), 5-24.

Wise, J. C., Sevcik, R. A., Morris, R. D., Lovett, M. W., Wolf, M., Kuhn, M.,Meisinger, B., &Schwanenflugel, P. (2010). The relationship between differentmeasures of oral reading fluency and reading comprehension in second-gradestudents who evidence different oral reading fluency difficulties.LanguageSpeech and Hearing Services in Schools, 41(3), 340-348.

Yassine, Saleh, (2004). The characteristics of the psychometric test of the reference of the reference in chemistry for the students of the first grade secondary scientific ability according to the classical and modern theories of measurement. Unpublished PhD thesis, Amman Arab University: Jordan.

Yurdabakan, I. (2012). Primary School Students' Attitudes Towards Computer Based testing and Assessment in Turkey. Turkish Online Journal of Distance Education, 13 (12), 177-188.