On the Impact of Illustrated Assessment Tool on Paragraph Writing of High School Graduates of Qom, Iran

Esmaeil Bagheridoust, PhD (Corresponding author)
Assistant Professor and Researcher
Islamic Azad University, South Tehran Branch
1st floor, No. 156, Khajeabdollah Ansary Ave., Shariati Street, 16616,Tehran, Iran
Tel.:98-21-228-611-86 E-mail: esmaeilbagheridoust@gmail.com

Zahra Husseini, MA
EFL/ESL Instructor
Islamic Azad University, South Tehran Branch
Avayesh Publications, No. 18, 12 Imami Alley, Basij Street, Qom, Iran, Postal Code 37158-13418
E-mail: zahra2007hhh@gmail.com

Received: March 14, 2011 Accepted: March 30, 2011 doi:10.5539/elt.v4n3p101

Abstract

Writing as one important skill in language proficiency demands validity, hence high schools are real places in which valid results are needed for high-stake decisions. Unrealistic and non-viable tests result in improper and invalid interpretation and use. Illustrations without any written research have proved their effectiveness in whatsoever context they are used. They also have demonstrated their effectiveness in language teaching context. The present research tries to investigate the efficacy of illustrations as an assessment tool in improving the English paragraph writing among high school graduates. Participants in two intact groups are offered a writing test with and without illustrations to identify if illustrations affect writing test results. Via SPSS software results were compared and illustrations proved to be effective in writing test results.

Keywords: Illustrated assessment tool, Evaluation, Illustrations, Paragraph writing, Assessing and testing illustrations and visuals, Rubrics to assess writing

1. Introduction

Learning has been an inseparable part of human beings. The truth is that, for teaching and learning processes of a sort, there should be some conduits for both teachers and learners to find out more about the quantity and quality of learning. Hence, teachers ought to figure out if any learning has taken place or not. To this very end of teaching, assessment has always been attached to provide teachers with cogent evidence that what they have done really worked. Learning assessment is also done by the learners to become aware of their progress, knowledge and needs for a specific goal, or by teachers to know about the strengths and weaknesses of their students and their teaching capabilities, or by the governmental system of education to diagnose the success or failure of a curriculum and to decide about curriculum revision, or by employers to find the most suitable person for their purpose.

Traditional language learning assessment was used for a long period to test the amount of learning in all skills of language teaching (listening, speaking, reading, writing, pronunciation, grammar, etc.). It consisted of a set of written tests of finding native equivalents for foreign words and sentences. Written tests were very popular; even aural/oral skills of language have been tested through written tests in schools (Heaton, 1990). Therefore it is not surprising to say that if a person has reading or writing problems and is a proficient listener or speaker, his/her scores on the tests did not demonstrate his/her real capabilities in listening and speaking skills in language being learned.

Anchored in the findings of researchers in the field of language testing such as Horne (2007), Raimes (1983), Madsen (1983), Heaton (1990) and Allen (1983), different approaches were designed to assess the overall learning in EFL students/learners. A best way that is being agreed upon was to assess what learners could do in a language and not what they knew about that language (Weir: 2005, Richards & Reynandya :2002, Richards & Rogers: 2002). Therefore introducing an alternative method for assessment seemed inevitable. In line with the new trends in psychology which found its place in the society and consequently in learning, a temptation appeared to present more communicative and less stressful tests to learners and adapt more humanistic approaches in the field of assessment, evaluation and testing. It was believed that even the process of assessment could facilitate, motivate and deepen learning (Weir: 2005, Richards & Reynandya :2002, Richards & Rogers: 2002). Some shortcomings of testing and tests were impossible to be solved completely but improvements were welcomed. If various demands of validity, reliability and practicality were believed to be necessary in written tests, improvements to these kinds of tests could be followed to increase validity, reliability and practicality even more.

Some scholars such as Raimes (1983), Madsen (1983) and Heaton (1990) have discussed the efficiency of illustrations in teaching and testing of different skills in second/foreign language. Nowadays illustrations are the main part of more textbooks and web pages without which it seems that something is missing. They are also the first attraction in any written context based on which a general judgment is made prior to any reading is done, may be even before the topic is read. FCAT writing which was established in Florida University and now is widely used for teaching writing is using illustrations in teaching and testing of writing (FCAT, 2007).

The researchers have decided to extend the use of illustrations to tests of writing among intermediate students. The difference is that in the present research illustrations would not be used in a way to limit the creativity of students by presenting them the picture of a scene to describe; or a series of pictures about which students should tell a story. The aim is to present students with some illustrations that activate the body of knowledge that students have about language to enable them to have a more trustable evaluation of their knowledge about language.

2. Research question

Based on the objectives of the study the researchers' question is:

"Does the use of illustrations in assessment of paragraph writing among EFL learners bring about any variation in their performance in the test?"

3. Statement of the hypothesis

Concerning the foresaid research question; the following null hypothesis is formulated;

"The use of illustrations in assessment of paragraph writing among EFL learners does not reliably bring about any variation in their performance in the test."

4. Theoretical background and purpose

Teaching and testing are the yin-and-yang of each other that investigation of one without dealing with the other seems quite unrealistic. J. B. Heaton (1990) asserts that:"Both testing and teaching are so closely interrelated that it is virtually impossible to work in either field without being constantly concerned with the other. Tests may be constructed primarily as devices to reinforce learning and to motivate the student or primarily as a means of assessing the student's performance in the language." (p. 5)

Anywhere where some teachings occur, there must be some kind of assessment to make teachers and administrators sure about the success or failure of the teaching course goals. Using some kind of test is the most convenient way to give administrators some idea about the success or failure of the teaching course due to the abilities of the teacher in teaching the course syllabus or the abilities of the students/learners in learning what they are expected to learn. The goal of the test is an important determinant (Weir: 2005, Brindley: 2001, Bachman & Palmer: 2000, Bachman: 1995, Heaton: 1990, Madsen: 1983).

A test of language tries to find out what learners can do with language in real life to achieve some goals. Approaching the issue from this viewpoint, testing, like teaching, is influenced greatly by psychology. Dependent on how teaching is viewed, how the learner is defined and how the process of learning is analyzed, a test is supposed to measure and evaluate the degree of learning achievement. (Huerta-Macias: 2002, Genesee: 2001). Tests can be divided into classroom tests and external examinations, as Heaton (1990) and Farhady et al (1994) believe. These tests are different from each other as far as the purpose of the test is concerned. The function of class tests is to locate the areas of difficulties an individual student or a group of students have in learning the course materials. The second objective of a classroom test is to evaluate the effectiveness of his/her method of teaching and materials s/he uses in the class. In these cases, some types of tests are needed to measure the capabilities of the class as fairly as possible, not to set up trap for the students to prove them that they have not performed as they should. A well-constructed classroom test gives the students the opportunity to find out about their ability to perform certain tasks in the target language. On the other hand, testing for assessment aims to distinguish the amount of learning in a target group. In this trend of testing, the way in which this learning has taken place is not the matter of concern. Whatever the goal of the test is, the test should be useful as L. F. Bachman and A. S. Palmer (2000) suggest: "The most important consideration in designing and developing a language test is the use for which it is intended, so that the most important quality of a test is its usefulness. This may seem so obvious that it need not be stated. But what makes a test useful? How so we know if a test will be useful before we use it? Or if it has been useful after we have used it?" (p. 17)

Brindly (2002) differentiates among assessment, evaluation and testing. Assessment refers to collecting data on the language ability of a learner or his/her achievement in a language course. Evaluation, on the other hand, refers to the overall performance of the students in a language program and has nothing to do with what individuals do. In this case, testing is a sub-component of assessment. Assessment is of two kinds: first, proficiency assessment that measures the language skill of the learner independent from any course of learning they have passed. Second, assessment of achievement that aims to measure what an individual has learned in a particular course of study. If an assessment is administered during the learning process to help the teacher modify the instruction, it is a formative assessment, but if it is performed at the end of the course to find out the cumulative course outcomes, it is a summative assessment. If the results of the test are interpreted for each student in comparison with other students, it is a norm-referenced test, but if these results are interpreted in comparison with a standard test results, it is a criterion-referenced test (Farhady et al: 1994).

To make the long story short, alternative assessment, if done with particular attention and with definite touch stones, will serve learners and teachers more than the standard assessments. The benefits of alternative assessment should not cause the teachers to ignore or underscore standard assessment. When there is a need to know exactly the amount of knowledge one has in a particular language component or when there is a huge group of candidates about which fast and reliable and valid decisions should be made in a short time, standard tests are the best choice.

Testing is expected to serve as a complementary to teaching. Teaching is a controversial issue and therefore it is impossible to prescribe a single best method for teaching in general and for writing in particular. Writing is chosen to be studied because according to several scholars it is regarded to be the last and the most important and the ultimate skill a language learner can acquire (Urquhart & McIver: 2005, Reid: 2001, Jordan: 1990, Raimes: 1983). It is used as a means to measure other language skills. Writing proficiency in a foreign language is possibly more useful than speaking or listening proficiency in that language because the place and time for a communication to occur is not provided very easily (Reid:2001). Things are not the same for writing skill; one does not need to travel to the country in which a foreign language is spoken. Everybody has the chance to start a written communication with people that are hundreds of miles far from the writer. New communication tools such as e-mails, blogs, and web pages have increased the emphasis on the writing ability.

An illustration can be any form of graphics that is used to attract attention to a specific point more effectively. Any type of photograph, painting, drawing calligraphy, chart, diagram, map or other works of art can be examples of illustrations in books. Illustrations can serve to decorate a story, poem or piece of textual information such as a newspaper article or magazine or school textbooks. They often provide a visual representation of something described in the text to present some additional message. Graphic design is the process and art of combining text and graphics. It is useful in communicating an effective message in the design of logos, graphics, brochures, newsletters, posters, signs, and any other type of visual communication. In today's world graphic designers often use computer software to achieve their goals. J. R. Levin and R. E. Mayer (1993) believe: …In countless laboratory learning contexts, a clear finding has emerged: Pictures are remembered far better, NS Fr longer, than are their verbal counterparts. …. This conclusion derives from experimental research paradigms that have incorporated simple, controlled, and often contrived, learning materials (typically, lists of unrelated objects/nouns or object/noun pairs). (p. 96)

Words and pictures serve to communicate information in textbooks. The society prefers words to illustrations in textbooks, but Levin & Mayer (1993) have explained in detail their own experiments and that of other researchers that claim the fact that researches are on the illustrations side and stress the positive impact of illustrations in textbooks. Pictures have usually been remembered longer than texts by the learners. If illustrations accompany texts, they would be remembered longer than the texts that lack illustrations. In the following come several reasons to explain why illustrations facilitate learning:

1) They attract students' attention to the most important part of the learning materials.

2) They summarize wordy explanations into a simple picture.

3) They provide a pictorial representation of a text and therefore reinforce it in the minds of learners.

4) They can simplify difficult and complicated data.

5) Sometimes unfamiliar information can be transformed to some memorable forms and improve learning in this way.

Learners' attitudes towards learning situations and learning materials are accepted as an important part of any learning situation (Richards & Rodgers: 2002, Brown: 2007). Illustrations can be useful in providing this positive effect almost in any learning context. When they were used in books they made books more attractive and appealing and publishers used them in various shapes and types to guarantee their market (Woodward: 1993). It did not last long when teachers and instructors recognized the importance of illustrations in facilitation of learning among students and ease of teaching among teachers and as a result illustrations interred into learning context (Brown: 2007, Mrjia et al: 1992, Larsen-Freeman: 1986, Levin & Hughey: 1985, Raimes: 1983). Publishers were really influential in expansion of illustration in academic books, especially when the tendency of competition among publishers to sell more books was evident (Woodward: 1993). The positive impact of illustrations paved the way for illustrations to be used in assessment of language skills and sub-skills (Madsen: 1983, Raimes: 1983). Madsen (1983) introduces various techniques to use visuals in test of listening comprehension. He also emphasizes that using visual for this goal is particularly useful for children and beginners.

5. Method

5.1. Participants

The population of the research is all students of schools in Qom, the most religious city in Iran, who were studying in pre-university schools hence all were high school graduates. The total number of the population was 8,187 who were studying in different types of pre-university schools such as Isaargaraan (schools for ex-soldiers), Tizhoshaan (schools for talented students), Shahid (schools for children of Martyrs), Moshaarekate Mardomi (schools to which people donate money), Ordinary State Schools, Nemooneh Dolati Schools (state schools that provide extra educational facilities for students and parents are to pay for those services), Private Schools, Distance Learning Schools and Schools that are related to some state organs. The researchers made up their mind to give the test to volunteers. After an interview with the authorities who were in charge of exam administration it was realized that the best way to access the most random population was to refer to the special exam centers that were limited in number and in which all high school students from all over Qom were preset. There were twelve exam centers in Qom of which six centers were specified for males and six centers for females. Qom consists of four educational districts and two little towns called Kahak and Jafar-abad. Each district and town had nominated two schools as centers for final exams of pre-university students: one for males and one for females. All of the centers were visited in six days. They were present in the exam centers about 30 minutes before each exam (exams started every day at 10:00 a.m. Before their exam they invited students not to leave immediately after the exam and see her before they leave. It was tried not to leave any student uninvited even when in some cases some students were late for exam, they received the same invitation.

There was no control on the absence or presence of some students in time of invitation. It can be claimed that all of the research population received the invitation. All those students who took part in the conversation with the researchers were asked them to be present in a theater saloon of Qom University on June 26, 2010 (Tir 5, 1389) at 3:30 p.m. They were assured that if they took part in that research they received 4 hours free English writing course and they would receive a complete guide to their writing strengths and weaknesses. One hundred eighty students were registered and their addresses were collected for later interactions. Only 128 students were present in the first presentation day of the test. And in the second day 17 more students were absent. After collecting the papers, 7 papers were regarded as unacceptable as they had written irrelevant materials or had cheated. As a result 104 papers were accepted and graded. The age of the participants was between 17- 19 years, 68 were male and 36 were females. Because participants were left free not to write their real names if they did not like to, it was almost impossible to separate male and female papers. The research did not aim to differentiate between sexes therefore there was no attempt to do so. All of the participants had finished pre-university course and were preparing themselves for final exam. The background knowledge of the participants was not controlled and they took part in the research with different or similar backgrounds.

5.2. Apparatus

Paper exams are common widely used equipments in most learning situations to assess students learning. They are also the main and the most important vehicle of assessment in Iranian schools. Therefore it was decided to use paper tests in the present research. Writing can be assessed in different forms such as substitution drills, fill in the blanks, controlled writing and free writing. Free writing style of testing was not innovative in Iranian high schools. The only difference was the illustrations that were added to the second test to trigger ideas in the students' minds. If they have no idea about what to write and focus only on writing which was the ultimate goal of the test.

6. The design of the study

The present research follows a cross-sectional design. Using pencil and paper makes the mode of survey administration personal (face-to-face). According to Fink & Kosecoff (1985), personal administration generally yields highest cooperation and lowest refusal rates and makes longer and more complex interviews possible. It also provides high response quality which is resulted from the interviewer presence in the survey. Taking these advantages into consideration in comparison with the disadvantage of this type of administration (including costly mode, longer data collection period and interviewer concerns), it would be acceptable to administer a face-to face type of data gathering. The design is quasi-experimental because subjects are not grouped. This is a kind of one-group pretest-posttest design. It is subject to such threats to validity as history (events intervening between pretest and posttest), maturation (changes in the subjects that would have occurred anyway), regression toward the mean (the tendency of extremes to revert toward averages) and testing (the learning effect on the posttest of having taken the pretest). The nature of the survey made this type of design inevitable. To reduce the above-mentioned undesirable factors some steps are taken: two very common issues are selected as topics of writing about which much is discussed in the mass media, schools and society to reduce the effect of history (these topics are also mentioned in the pre-university course of students); two weeks interval cannot have a noticeable influence on maturation. This study attempts to investigate a causal relationship via a quantitative research design.

7. Procedure

Only some of test techniques can be performed with illustrations such as free writing about a topic or controlled writing. Urquhart & McIver (2005), Raimes (1983), Madsen, (1983) stress on different types of free writing as a way to test writing. This method of testing was very familiar for Iranian students. The purpose of the present research was to compare a single test with and without illustrations. Therefore choice should be given to the students to act freely in each test to make comparison possible and meaningful. Ten high school teachers were interviewed to give their ideas about the possible suitable topics. Their diversity of ideas and lack of a firm decision about some single topics pushed the researchers towards the pre-university English language teaching book (Anaani Sarab & Samimi: 2009). Using high school book could help to improve the content validity of the test. All lesson topics were listed and "protection from earthquake" and "the relationship between wealth and happiness" were chosen. Because these topics were talked about greatly in mass media and it was believed that there was a sufficient amount of knowledge available in the mind of participants about the topics. Moreover these topics seemed the most suitable ones for which lots of illustrations could be found. Students were free to choose between topics.

Similar to formal traditional tests of compositions a single topic was written on top of the exam sheet with a traditional slot in which students were required to write their names. In the other version of the test the same topics were presented in the same way. This time some illustrations decorated the paper. These illustrations were chosen in a way to activate the students minds regarding the topics, to be simple and easily understandable, not to be complicated to confuse students and engage their minds too much and to follow the forms of sketches to be easily xeroxed (as xeroxing is the most and the main way of preparing exam tests in Iran). Therefore illustrations should have as few small details as possible.

It was attempted to provide a situation similar to the real examination sessions and behave students as if they were in an authentic exam session and the topics were read aloud for students and they were asked if they had any problem in understanding the topics or if they had any ambiguity in what they should do. Students were asked to write about one of the mentioned topics about 100 words in 45 minutes. All of the questions of students were answered to guarantee at least test face validity by preventing any misunderstanding of students in the test. They were encouraged to write their name for later comparison with their second paper but this was not told to them. They were only told to write their names to be informed about their grades. They were also told that it was alright to choose an imaginary name if they did not like to write their real names but it was emphasized that they had better to remember that name for their next session. They were not informed about the real reason lest it affect their next exam and made them prepare themselves for the next exam. Eight colleagues of the researchers helped them to conduct the test. In the second presentation of the test the same test was presented with the same characteristics. Students were asked to write their names on the top of papers. This name real or imaginary should be the same as the previous one to make it possible for comparison. Participants were asked to choose the same topics that they had written about in the previous session. The test was administered in a similar fashion as before and papers gathered for later studies.

One of the most complicated aspects of the present research was to choose proper criteria (writing rubric and scale) to grade the writings of the students. Among various rubrics (GETWORKSHEETS, KIMMELSKORNER, FCAT, IMET. FAERIEKEEPER, NESTERSTEACHINGBLOG) the FCAT writing criteria was chosen (retrieved from http://jmsenglish.com). It was difficult to decide which one to choose and a suitable justification was necessary for rejecting the others. However as one was to be selected on the one hand and FCAT was derived from a pretty well-known university, the researchers finally chose FCAT from among others including the above-mentioned ones. There were some difficulties in choosing such criteria for grading the papers, for example; they were suitable for larger pieces of writing such as articles. They were not applicable to writing abilities of the lower-intermediate students who were not proficient writers. The scales were from 1 to 6 and it could not discriminate properly among certain types of writings which were very close to each other. Two raters (to become sure about the inter-rater consistency in order to improve the objectivity of the test results) corrected the papers according to FCAT Rubric and gave them credits.

8. Results and discussion

The first step in the analysis of null hypotheses is to become sure that the FCAT rubrics provide a valid measure of the test. Horne (2003) presents a review of the reliability of the writing scores that were corrected with these rubrics and proves that these rubrics provide reliable results. The researcher also calculated the reliability of this rubric for this specific experiment through paired t-test formula by assuming the scores of the first rater as one test and the scores obtained by the other rater as another test. Hence the mean difference would be equal to 0.0769 for the un-illustrated test scores and -0.0385 for the illustrated test scores. The p-value was smaller than 0.05 in both calculations. These mean differences are close to zero and this indicates the measurement tool is highly reliable. Inter-rater reliability is calculated through test retest method that is 0.994 for the un-illustrated test and 0.099 for the illustrated one. This can be also an indication of rubrics reliability.

The statisticians, with whom the researcher interviewed, believed that it was impossible to calculate reliability in human sciences. They believed that in human sciences only assimilations can be made to the paradigms of statistics. In the present research, therefore, it should be assumed that each student knew that writing consisted of Focus, Organization, Support and Convention. Then each component will be accepted as a sub-test and consequently the Cronbach's Alpha method can be used to prove the reliability of the test that is used in this research in illustrated and un-illustrated form. This reliability was 0.951 for the un-illustrated test and 0.0925 in the illustrated one.

A test should be reliable to be valid (Bachman 1995). Now that the reliability of the test is proved, its validity will be discussed. The test form seemed valid to the researcher since it was presented in the form which was prescribed by some scholars such as Farhady et al (1994) & Raimes (1983) for free writing tests. It also contained content validity because it focused on test of writing proficiency for which it was designed. It was also based on the content that was studied in pre-university book (Anaani Sarab & Samimi 2009). Construct validity was also apparent in the present test because it followed the theories of teaching writing which was explained in works of Raimes (1983) and Urquhart & McIver (2005): in both cases free writing about a stated topic were accepted as a valid way to evaluate writing. The next step was to decide on how to insert illustrations in the test. Centering the illustrations was very eye-catching but the problem was the blank area in which students were to write their papers. It was suitable to place illustrations in the left or right side of the paper. Most of the people are right-handed so if illustrations were placed on the right side they were probably more eye-catching.

9. Data analysis of the first experiment

In the un-illustrated test standard deviation from the mean is almost 6.2 that is almost normal when the number of candidates are considered. Results suggest the greatest standard deviation is in Conventions and the least is in Focus. The difference between the highest score and the lowest scores is very great: it is 6 in each section and 24 in the total scores that is equal to the total scores one could obtain. It means that there were some students who received the total scores and some who received none. For further investigation each group divided into three parts and the sum, mean, median, mode, range and standard deviation of their scores calculated separately. The number of students was 104 and therefore the first and the last students' scores were omitted in order to have three equal groups. Group 1 represents the upper third section of the total group who took part in the un-illustrated test. Group 2 and group 3 represent the middle and lower thirds in the group respectively.

10. Data analysis of the second experiment

Since the mean is 12.74, the distribution of sub-scores (Focus, Organization, Support and Conventions) is more diverse than those in the previous test. Standard deviation from the mean is almost 5.4 that is smaller than the previous exam representing a more homogenous test result. Results suggest the greatest the standard deviation is in support and the least is in focus. The difference between the highest score and the lowest one is 5 in each section and 20 in total scores. It means that nobody tried to leave the answer sheet blank. For further investigation each group divided into three parts and the sum, mean, median, mode, range and standard deviation of their scores calculated separately as it was performed in the first experiment. These scores are not expected to be compared to each other but they are provided for minute investigations in differences between the first and the second test. The greatest amount of fluctuations is apparent in Group 1 where range is 10 and standard deviation is almost 3.7. Group 2 represents the least fluctuations where range is 9.5 and standard deviation is almost 2.6. Although the standard deviation of scores in Group 3 is not the greatest in comparison to two other groups but this group possesses the greatest range scores that is 16. To conclude, based on these two sets of scores from the two administrations, researchers try to present a thorough analysis of the data in order to investigate the effect of inserting illustrations in the writing tests of high school students in Qom. These data can illuminate the authenticity of the null hypotheses that were present in the beginning chapter of the present research.

11. Analysis of the experiment results

In order to investigate the data it is better to present the analysis of the test scores of the two tests in the same page for a more comprehensible comparison. As shown in Table I. below, standard error of mean is decreased in the illustrated test. The variance of scores from the mean is greater in the un-illustrated test. Skewness of the scores in both tests is positive and this signifies that most of the scores in both tests are greater than mean and this positive skewness in the illustrated test is higher. Kurtosis suggests the dispersion of scores in comparison with normal distribution: it is negative in both tests and this proves the greater dispersion of scores in these tests in comparison with normal distribution. This dispersion is greater in the illustrated test and hence signifies a higher discrimination among students. In both test Std. Error of Skewness and Std. Error of Kurtosis are greater than +2 and it means that the distribution of scores is not normal.

Here it is suitable to re-state the null hypothesis of the research:

"The use of illustrations in assessment of paragraph writing among EFL learners does not reliably bring about any variation in their performance in the test."

If the mean of the un-illustrated test is $\mu 1$ and the mean of the illustrated test is $\mu 2$, the null hypothesis says:

Ho: $\mu 1 = \mu 2$

and if H1: $\mu 1 \neq \mu 2$ the null hypothesis will be rejected.

To test this hypothesis paired sample t-test is used. This formula has the confidence interval that is lower than 0.05 (sig) and investigates the effect of an intervening factor among two groups.

12. Review of the null-hypothesis

In this part the researchers attempted to calculate the impact of illustrations as an intervention in the illustrated test in comparison with the un-illustrated. Using SPSS software to calculate t-test the correlation between two tests is: r = 0.78 with a significance (p-value) lower than 0.05 that approves a confidence interval of 0.95 percent. This high correlation can be a proof of the efficacy of measurement tool. What should be noticed is the point that scores of the students in the first test highly correlates with their scores in the second test. This correlation value indicates a meaningful difference between scores of the un-illustrated test and scores of the illustrated one. SPSS software presents paired mean difference of the two tests as follows: $\mu 2$-$\mu 1$ = 2.33 and therefore $\mu 1 \neq \mu 2$ and therefore Ho: $\mu 1 = \mu 2$ will be rejected.

13. Final discussion

It was discussed in detail that mean of total scores have improved the illustrated test in comparison with the un-illustrated one. Therefore the null hypotheses will be rejected, i.e. the use of illustrations in assessment of paragraph writing among EFL learners brings about some variations in their performance in the test. For a thorough analysis of the experiment, the researchers divided the participants into 3 groups. To provide three homogenous groups one participant from the beginning and one from the end of the list were omitted. The remaining 102 participants were divided into three groups, each consist of 34 members: the upper third, the middle third and the lower third. Descriptive statistics for all groups are illustrated below.

Standard error of mean is increased in the illustrated test for all groups. Skewness of the scores in the un-illustrated test is positive for upper and middle groups and negative for the lower group and this signifies that most of the scores are greater than mean for two upper thirds of the sample population. Skewness results value is positive for middle and lower groups and negative for the upper group in the illustrated form of the test. In upper and middle groups skewness has decreased to provide evidence for scores gathering near the mean while a shift from negative skewness towards a positive one in the lower group claims the improvement of scores. Kurtosis suggests the dispersion of scores in comparison with normal distribution: it is negative in both upper and lower groups' un-illustrated tests scores and positive in the middle group, tests and this proves the greater dispersion of scores at the two ends of the population and lower dispersion in the middle group in comparison with normal distribution. In the illustrated test, however, kurtosis suggests greater dispersion of scores in two upper groups and smaller dispersion in the lower group. In both test Std. Error of Skewness and Std. Error of Kurtosis is smaller than +2 and it means that the distribution of scores is not normal.

Following are the discussion of the comparison of correlations of the groups. In the upper third group (table II. Describes the case in more detail below) r = 0.825, and for the middle group it is r = 0.344 and for the lower group correlation equals to -0.065. It indicates a positive and strong relationship among more knowledgeable students and a positive and relatively strong relationship among average students. The negative correlation value in the lower group that is very close to zero proves that the use of illustrations in paragraph writing test of less knowledgeable students does not bring about any meaningful change in their performance in the test.

An accurate study in the above tables will show that it is possible to improve the performance of students (strong, average or weak) in their paragraph writing tests; these illustrations are more beneficial to weak students, average students also benefit from illustrated writing tests and strong students also will find illustrated tests useful although its advantages for them is small.

References

Anani Sarab, M. R., & Samimi, D. (2009). English for Pre-University (1&2). Tehran: Sherkate Chaap va Nashre Ketaab-haaye darsi-ye Iran

Bachman, L. F.(1995). Fundamental Considerations in Language Testing (3rd ed.). Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2000). Language Testing in Practice (3rd ed.). Oxford: Oxford University Press.

Brindley, G. (2001) Assessment. In R. Carter & D. Nunnan (eds.) The Cambridge guide to teaching English to speakers of other languages. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511667206.021, http://dx.doi.org/10.1017/CBO9780511667206.021

Britton, B. K., & Woodward, A. & Binkley, M. (eds.). (1993). Learning from textbooks. New Jersey: Lawrence Erlbaum Associates, Inc.

Brown, H. D. (2007). Principles of Language Learning and Teaching (5rd ed.). USA: Longman.

Denman, B. R. (2000). In contact (1): Beginning, second edition. New York: Longman.

Farhady, H., Jafarpoor, A., & Birjandi, P. (1994). Language testing skills: from theory to practice. Tehran: SAMT.

Fink, A., & Kosecoff, J. (1985). How to Conduct Surveys: A Step-by-step Guide. Beverly Hills, CA: Sage.

Genesee, F. (2001) Evaluation. In R. Carter & D. Nunnan (eds.) The Cambridge guide to teaching English to speakers of other languages. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511667206.022, http://dx.doi.org/10.1017/CBO9780511667206.022

Heaton, J. B. (1990). Writing English Language Tests (3rd ed.). NewYork: Longman Inc.

Horne J. (2003). Reliability and Validity of FCAT: Assessment & Accountability Briefing Book: School Accountability Certification FCAT Teacher Tests.

Huerta-Macias, A. (2002) Alternative assessment. In J. C. Richards & W. A. Reynandya (eds.) Methodology in language teaching. New York: Cambridge University Press. doi:10.1017/CBO9780511667190.048, http://dx.doi.org/10.1017/CBO9780511667190.048

Jordan, R. R. (1990). Academic writing course. London: Collins ELT.

Larsen-Freeman, D. (1986). Techniques and principles in language teaching. Hong Kong: Oxford University Press.

Levin, L. & Hughey, L. S. (1985). Changing times: towards an integrated approach to reading. Englewood Cliffs, New Jercy: Prentice-Hall, Inc.

Levin, J. R. & Mayer R. E. (1993) Understanding illustrations in text. In B. K. Britton & A. Woodward & M. Binkley (eds.) Learning from textbooks. New Jersey: Lawrence Erlbaum Associates, Inc.

Madsen, H. S. (1983). Techniques in testing. Hong Kong: Oxford University Press.

Markstein, L. & Hirasawa, L. (1977). Expanding reading skills: advanced. Massachusetts: Newbury House Publishers, Inc.

Mrjia, E. A., & Xiao, M. K., & Pasternak, L. (1992). American picture show: a cultural reader. Englewood Cliffs, New Jercy: Prentice-Hall, Inc.

Raimes, A. (1983). Techniques in teaching writing. Hong Kong: Oxford University Press.

Richards, J. C., & Rogers T. S. (2002). Approaches and Methods in language teaching (2nd ed.). New York: Cambridge University Press. doi:10.1017/CBO9780511667190, http://dx.doi.org/10.1017/CBO9780511667190

Reid, J. (2001) Writing. In R. Carter & D. Nunan (eds.) The Cambridge guide to teaching English to speakers of other languages. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511667206.005, http://dx.doi.org/10.1017/CBO9780511667206.005

Urquhart, V. & McIver, M. (2005). Teaching Writing in the Content Areas. New York: ASCD.

Weir, C. J. (2005). Language Testing and Validation. New York: Palgrave Mcmillan.

Woodward, A. (1993) Do illustrations serve an instructional purpose in U.S. textbooks?. In B. K. Britton & A. Woodward & M. Binkley (eds.) Learning from textbooks. New Jersey: Lawrence Erlbaum Associates, Inc.

Yule, G. (1985). The story of language. New York: Cambridge University Press. [Online] Availale: http://www.fldoe.org/Strategic_Plan (January 2, 2011)

Retrieved January 2, 2011from http://www.getworksheets.com/samples/rubrics/writing.html

Retrieved January 2, 2011, from http://www.kimmelskorner.com/fcat_writing_advice.htm

Retrieved January 2, 2011, http://imet.csus.edu/imet6/bundy/classes/imet281/mylesson.html Retrieved January 2, 2011, http://faeriekeeper.net/criteria38.htm

Retrieved January 2, 2011, http://nestersteachingblog.wordpress.com/

Table1. Statistics of both experiments

		Total A	Total B
N	Valid	104	104
	Missing	0	0
Std. Error of Mean		.6058	.5281
Variance		38.165	29.005
Skewness		.225	.407
Std. Error of Skewness		.237	.237
Kurtosis		-.505	-.738
Std. Error of Kurtosis		.469	.469

Total A: Total scores of students in the un-illustrated writing test

Total B: Total scores of students in the illustrated writing test

Table2. Paired Samples Test of the three thirds

		Paired Differences							
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
					Lower	Upper			
Pair 1	Total B – Total A Upper Third	.9706	2.1068	.3613	.2355	1.7057	2.68	33	.011
Pair 1	Total B – Total A Middle Third	1.4559	2.4475	.4197	.6019	2.3099	3.46	33	.001
Pair 1	Total B – Total A Lower Third	4.2206	4.7197	.8094	2.5738	5.8674	5.21	33	.000